


June 2022

Psychometric Characteristics of Academic Language Discourse Analysis Tools

Courtney (Cici) Brianna Claar
University of South Florida

Follow this and additional works at: <https://digitalcommons.usf.edu/etd>

 Part of the [Arts and Humanities Commons](#), [Educational Assessment, Evaluation, and Research Commons](#), and the [Elementary Education and Teaching Commons](#)

Scholar Commons Citation

Claar, Courtney (Cici) Brianna, "Psychometric Characteristics of Academic Language Discourse Analysis Tools" (2022). *USF Tampa Graduate Theses and Dissertations*.
<https://digitalcommons.usf.edu/etd/9320>

This Ed. Specialist is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact scholarcommons@usf.edu.

Psychometric Characteristics of Academic Language Discourse Analysis Tools

by

Courtney (Cici) B. Claar

A thesis submitted in partial fulfillment
of the requirements for the degree of
Educational Specialist
with a concentration in School Psychology
Department of Educational and Psychological Studies
College of Education
University of South Florida

Major Professor: Trina Spencer, Ph.D., BCBA-D
Jose Castillo, Ph.D.
Yi-Jui Iva Chen, Ph.D.

Date of Approval:
June 22, 2022

Keywords: academic language, language sampling, language comprehension, language measures

Copyright © 2022, Courtney B. Claar

ACKNOWLEDGMENTS

I would first like to express my deepest gratitude to Dr. Trina Spencer, my mentor, academic advisor, committee chair, and friend. Your encouragement, dedication, leadership, patience, feedback, and example have been indispensable to my growth and development, not only as a scholar, but as a person. Additionally, this project would not have been possible without my defense committee, whose knowledge and expertise were invaluable assets. I would also like to acknowledge the Institute of Education Sciences (IES) for financing the study from which the data of interest were sourced.

I am grateful to the ALPS angels, my friends and research comrades who dedicated themselves to hours of data collection and tedious transcription. Thanks should also go to the researchers and administrative staff at the Rightpath Research and Innovation Center, especially those who helped with developing the *NLM* and *ELM Flowcharts* and those who scored the many thousands of language samples for this project. Thank you for motivating and inspiring me. Special thanks also goes out to individuals from the USF Educational Measurement and Research department, who put in considerable work in helping prepare data for this analysis.

Lastly, I would like to thank my family, especially my parents and grandparents, for their belief in me and continued investment in my education. Their support made it possible for me to reach these academic and professional goals, and I am deeply grateful.

TABLE OF CONTENTS

List of Tables	iii
List of Figures	iv
Abstract	v
Chapter One: Introduction	1
Academic Language.....	2
How Academic Language Influences Student Outcomes.....	5
Importance of Measuring Academic Language.....	7
Definitions of Key Terms	8
Chapter Two: Literature Review	10
Assessments of Spoken Academic Language.....	10
The Standardized Test Approach.....	10
Norm-Referenced Tests	11
Criterion-Referenced Tests	13
The Structural Assessment Approach.....	16
Structural Assessments of Narratives	19
Structural Assessments of Expositions	20
Strengths and Limitations of Current Approaches.....	22
Standardized Tests	22
Structural Assessments	24
Purpose of the Current Research Study	26
Importance of the Current Research Study.....	27
Chapter Three: Methods and Procedures.....	28
Participants	30
Research Team.....	31
Flowchart Materials and Standardized Procedures.....	32
Data Collection	32
Transcription and Coding	35
Overview of Data Analysis Strategy.....	35
Estimates of Validity.....	35
Estimates of Reliability.....	36
Interpreting Psychometric Properties of Productive Language Assessments	37

Chapter Four: Findings and Results.....	39
Descriptive Statistics.....	39
Differences across Grade Levels	41
Factor Structure of <i>NLM</i> and <i>ELM Flowcharts</i>	43
<i>ELM Flowchart</i> Model Specifications.....	45
<i>NLM Flowchart</i> Model Specifications.....	48
Reliability of <i>NLM</i> and <i>ELM Flowcharts</i>	49
Internal Consistency.....	49
Agreement between Raters	51
Correlations with Other Measures	54
Chapter Five: Discussion	55
Measurements of Reliability.....	57
Interrater Agreement.....	58
<i>Language Complexity</i> Subscale	58
<i>Narrative Structure</i> Subscale.....	58
<i>Passage Structure</i> Subscale	59
Internal Consistency.....	59
Measurements of Validity.....	60
<i>NLM Flowchart</i>	60
<i>ELM Flowchart</i>	62
Conclusions, Implications, and Future Directions.....	64
<i>NLM Flowchart</i>	64
<i>ELM Flowchart</i>	66
Support for Finding Alternatives to Item Deletion.....	67
Limitations of the Current Study	70
References.....	71
Appendices.....	81
Appendix A: <i>NLM Flowchart</i> (Front).....	82
Appendix B: <i>NLM Flowchart</i> (Back)	83
Appendix C: <i>ELM Flowchart</i> (Front).....	84
Appendix D: <i>ELM Flowchart</i> (Back)	85
Appendix E: Narrative Elicitation Script Example.....	86
Appendix F: Narrative Elicitation Script Example.....	87

LIST OF TABLES

Table 1:	Metrics Commonly Employed in Microstructural Analyses	18
Table 2:	Demographic Characteristics of Participants ($n = 1,040$).....	33
Table 3:	Five Phases of Data Collection and Scoring.....	34
Table 4:	Descriptive Statistics.....	42
Table 5:	Fits of Models that Test Different Conceptualizations of Narrative and Expository Academic Language	47
Table 6:	Factor Loading Analysis.....	50
Table 7:	Intercorrelation Estimates	51
Table 8:	Cronbach's Alpha Coefficients.....	51
Table 9:	Interrater Agreement and Cohen's Kappa Values	53
Table 10:	Correlations with WJ-TOL CALP Scores	54

LIST OF FIGURES

Figure 1: Conceptual Map of Academic Language via the NLM and ELM Flowcharts	29
Figure 2: Subscales and Items of the <i>NLM</i> and <i>ELM Flowcharts</i>	30
Figure 3: Proposed Factor Structure of Narrative and Expository Academic Language	38
Figure 4: <i>NLM</i> and <i>ELM Flowchart</i> Scores across Grade Levels.....	41

ABSTRACT

Academic language plays a key role in students' educational success, yet its development in primary grades is poorly understood and often neglected (Snow & Uccelli, 2008). Academic language skills may enhance overall academic performance if targeted early and intensively. However, current methods of assessment are not sufficient to understanding the construct well enough to develop evidence-based intervention strategies. This investigation examined the psychometric properties of two discourse analysis tools designed to directly measure students' comprehension and production of academic language. Academic language samples ($n = 7,887$) from a previous cohort-design study ($n = 1,040$; Kindergarten through third grade participants) were scored using the *Narrative Language Measure (NLM) Flowchart* and the *Expository Language Measure (ELM) Flowchart*. A confirmatory factor analysis was used to test two-factor models for both flowcharts. The total scores and subscale scores of the *NLM Flowchart* demonstrated moderate to strong interrater reliability, moderate convergent validity, and approximate fit with the proposed model (generation $\chi^2(46) = 743.85, p < .001$, SRMR = .06, RMSEA = .08, CFI = .88, and TLI = .86; retell $\chi^2(46) = 784.80, p < .001$, SRMR = .05, RMSEA = .09, CFI = .91, and TLI = .90). One subscale (i.e., *Narrative Structure*) showed adequate internal consistency via Cronbach's alpha. This study found mixed evidence of interrater reliability for the *ELM Flowchart*, with weak agreement on one subscale (i.e., *Passage Structure*) and substantial to strong agreement on the other (i.e., *Language Complexity*). The *ELM Flowchart* demonstrated moderate convergent validity, but neither subscale reached

acceptable levels of internal consistency via Cronbach's alpha. The appropriateness of using reflective indicator tools to evaluate constructs that may be better suited to a formative model is discussed. Other implications of the findings also are discussed.

CHAPTER ONE: INTRODUCTION

According to the most recent National Assessment of Educational Progress (NAEP), students have made alarmingly little progress in reading performance over the past 30 years and are currently on a downward trajectory (National Center for Education Statistics, 2019). It is now well understood that two critical skill repertoires form the groundwork of reading comprehension: word recognition and language comprehension (Gough & Turner, 1986; Hoover & Gough, 1990). Results of several meta-analytic reviews confirm that the application of theoretically grounded, evidence-based interventions targeting word reading skills contribute to positive, long-term gains in reading performance (Suggate, 2016; Goodwin & Ahn, 2010). However, the science of language comprehension is considerably less understood. Reading researchers suggest that unique variance in reading comprehension may be mediated in part by higher-order cognitive skills such as inference making, working memory, recall of previously learned information, and comprehension monitoring (Cain et al., 2001; Cain et al., 2004). However, knowledge of the constellation of linguistic skills necessary for comprehension, often referred to as academic language, is only beginning to develop among reading scholars. To advance the science of language comprehension, a deeper understanding of the word-, sentence-, and discourse-level patterns of language is needed (Adlof & Hogan, 2019; Phillips Galloway et al., 2020).

Scholars suggest that discovering how to teach academically relevant language skills is the next vitally important frontier of reading research (Cervetti et al., 2020; Phillips Galloway et

al., 2020; Schleppegrell, 2001; Snow & Uccelli, 2008). Theoretical and empirical investigations into the phenomenon of academic language should guide the development of curricula and interventions targeting this critical skill set. However, existent methods of assessing academic language are insufficient to the task of understanding its often nuanced and varied features. Theoretically grounded, valid, and feasible discourse analysis tools are needed to achieve the deeper level of understanding required to develop effective academic language instruction. Moreover, for academic language to be promoted intensely in primary grades, tools that inform academic language instruction and monitoring in schools are sorely needed.

Academic Language

Not all language is equally important to student learning. Language both shapes and is shaped by the social context in which it is used, making it a particularly dynamic, multidimensional, and context-specific construct (Halliday, 1993). Academic language is situated within the sociocultural context of school settings, enacting specialized communicative forms and functions. While individuals use everyday language ubiquitously to navigate social situations, academic language is specifically designed to convey complexity, higher order thinking, and abstraction (Zwiers, 2013). Colloquially referred to as the “language of schooling”, academic language is a more formal, specialized language used to communicate abstract, complex and technical ideas associated with academic disciplines (Schleppegrell, 2001). It is a way of communicating that fosters critical thinking and allows for a deeper, more precise understanding of academic content. In essence, it is the language used to acquire and express knowledge.

A comprehensive definition of academic language has been elaborated in various ways (Snow & Uccelli, 2008). Indeed, academic language contains a wide range of discrete features that can be difficult to integrate into a single model of communication. In the early 1980's, Cummins first explicated the distinction between basic interpersonal communication skills (BICS) and *cognitive academic language proficiency* (CALP) (Cummins, 1979). Corpus analyses have since confirmed that academic texts contain distinct structural features that set them apart as a unique register of language (Biber & Conrad, 1999). Importantly, academic language can be produced orally (spoken text) or in writing (written text). Some literature indicates that written language may be more formal than oral language (Gottlieb & Ernst-Slavit, 2014), but there is little if any empirical research that verifies this supposition among primary grade students.

The majority of academic language interventions have focused exclusively on vocabulary (Gottlieb & Ernst-Slavit, 2014; Uccelli et al., 2015). However, the scope of the academic language construct extends well beyond vocabulary and includes a range of complex grammatical and discourse features (Schleppegrell, 2001). Lexical features, or vocabulary, refers to the individual words produced in a text, spoken or written. These are perhaps the easiest characteristics to identify in student language. Grammatical features deal with the ways in which words come together to form sentences. Academic texts contain a greater density of complex grammatical structures, such as subordinate clauses and elaborated noun phrases, than colloquial language (Biber et al., 2011). The term “discourse” refers to any unit of communication, spoken or written, that is longer than a sentence. Discourse features consist of the ways in which sentences are organized to make texts intelligible and coherent. Grammatical and discourse features have been largely overlooked in the literature concerning interventions to enhance

academic language skills (Uccelli et al., 2015). This is especially true for young learner populations (Snow & Uccelli, 2008).

Educational linguists have identified a number of cross-disciplinary lexical, grammatical, and discourse features of academic language (Uccelli et al., 2014; 2015). However, within the academic language register, narrative and expository texts vary considerably according to form (structure), function (purpose) and context of elicitation. These two sub-registers, or discourse types, serve different purposes and generally take on distinct forms in relation to their purpose.

Narrative texts convey experiences and emotions, translate historical information, dispatch cultural knowledge, and/or teach morals and lessons. The story is a “fundamental instrument of thought” that plays an essential organizing role in cognitive processing (Turner, 1996, p.5). Stories help us plan, predict, explain, and think rationally about the world around us. Thus, the narrative form is ubiquitous in human affairs and essential for students to master. Developmental trends in children’s narrative production and comprehension are now well documented and understood (Curenton & Justice, 2004; Curenton, 2011). Research indicates that, when provided with explicit knowledge of and experience with narratives, young children make positive gains not only in language comprehension, but also in listening, reading, and peer relations (Johnston, 2008). In the U.S., narratives are expected to be in a linear time-sequence and contain specific discourse components. Stein and Glenn (1979) defined essential story elements to include *a setting* with reference to a specific protagonist, either an *initiating event* or an *internal response to an event*, an *attempt* to attain a goal, and a *consequence* signifying whether or not the goal has been reached.

Expository texts serve to inform, teach, argue and persuade audiences (Nippold, 2014). While all expository texts share some common structural characteristics, such as beginning with

a main idea followed by key details, they take on different organizational forms according to their purpose. Researchers have identified five organizational structures commonly employed in expository texts, namely *sequence* (situates ideas along a timeline); *cause/effect* (causal relations between ideas); *comparison* (similarities and differences among ideas); *problem/solution*, and *description* (general or specific information about a topic) (Shanahan et al., 2010; Duke et al., 2011). A significant body of research indicates that explicit instruction of the purpose and format of each of these text structures improves reading comprehension for both older and younger students (Meyer et al., 2014; Williams et al., 2007).

Expositions are thought to be more complex than narratives. Compared with narratives, expository texts often contain higher densities of low-frequency, technical vocabulary words, content words (e.g., nouns rather than prepositions) and complex linguistic devices, such as nominalizations (turning a verb into a noun; e.g., *employ* - *employment*) and subordinations (linking ideas through embedded clauses, e.g., *the situation **that we find ourselves in***) (Lundine & McCauley, 2016; Schleppegrell, 2001; Snow, 2010). Researchers have pointed out that a noticeable slump in students' reading comprehension accompanies the precipitous introduction of expository texts in most fourth-grade reading curricula (Best et al., 2008). Hence, sufficient preparatory materials for understanding expositions are needed to support the literary demands of this type of discourse.

How Academic Language Influences Student Outcomes

In response to the Common Core State Standards (CCSS) Initiative established in 2010 (National Governors Association Center for Best Practices & Council of Chief State School Officers), most states have engaged in extensive reform efforts to raise academic standards for

U.S. students. The new standards are designed to be more rigorous and better aligned with the requisite skills for occupational and academic success in adulthood. One of the content areas that has broadened significantly under the CCSS is oral communication. For example, there is a stated expectation for kindergarten students to orally produce personal stories and retell simple stories with information including characters, settings, and main events. Additionally, the Common Core State Standards Initiative (2021) places a heightened emphasis on nonfiction texts, especially for students in primary grades.

Why this enhanced emphasis on oral academic language? Listening and speaking, which are oral language repertoires, are the necessary foundation of reading and writing (CCSS, 2010). In a large-scale longitudinal analysis of students from kindergarten through tenth grade, Foorman and colleagues (2017) found that word recognition and language together explained nearly 100% of the variance in reading comprehension. It is also well established that language skills in preschool predict reading comprehension in third and fourth-grade (NICHD Early Child Care Research Network, 2005; Storch & Whitehurst, 2002).

Structural features of academic language begin to emerge in children's oral communication as early as three years old (Scheele et al., 2012). Research confirms that children need to develop specialized language skills to become literate. Oral academic language matures through sustained exposure to rich vocabulary, linguistically stimulating home and school environments, and opportunities to engage in extended discourse, all of which relate to reading achievement (Dickinson & Tabors, 2002). The contribution language makes to reading performance increases as students progress into upper grades (Geva & Farnia, 2012). This finding is not exclusive to, but more pronounced for English Language Learners (ELLs) than for English-as-a-first language

(EL1) students. Hence, development of key oral academic language skills may be particularly important for the reading development of ELLs (Lesaux, 2006).

Unfortunately, precise explication of the academic language skills required to build students' reading is obscured in much of the literature. Some studies have emphasized vocabulary as a primary predictor for reading comprehension (Hutchinson et al., 2003; Proctor et al., 2005). However, many investigations link reading comprehension outcomes to a wider scope of language skills that includes grammatical and discourse features (Farnia & Gena, 2013; LaRusso et al., 2016; Scheele et al., 2012; Uccelli et al., 2015). It is generally agreed upon that variance in students' reading performance can be explained by various components of language, including vocabulary, grammar, and listening comprehension; however, the precise contribution made by each component is, as yet, unknown.

Importance of Measuring Academic Language

Measurement matters. In the field of education, high-quality measures enable educators, parents, students and others to understand how students perform and to observe development over time. Academic language skills emerging in early childhood are predictive of reading performance in late childhood and adolescence (Scheele et al., 2012; Uccelli et al., 2015). These malleable factors may enhance academic performance if targeted directly using evidence-based instruction. However, what educators do not measure and understand cannot be taught.

At current, most approaches to measuring academic language are not comprehensive in scope. Instead, it is common for researchers to analyze individual features or constellations of related features in isolation. For example, an abundance of child language research reports findings related to the grammatical complexity of various registers of students' oral and written

communication (Biber et al., 1999). Structural assessments of the lexical/grammatical features of student language commonly report metrics such as mean length of utterance (MLU), clausal density, and use of connectives as indicators of linguistic complexity (Cahill et al., 2020; Guo et al., 2021; Granados et al., 2021). In contrast, a variety of test instruments have been developed to measure academic vocabulary growth in an effort to enhance the language skills of ELL populations (Truckenmiller et al., 2019). Finally, an extensive array of test instruments have been developed that measure narrative language production, another vital feature of academic language (Pesco et al., 2017). However meaningful these individual measures may be, the construct of academic language extends beyond any of the isolated features they assess, since academic language is both dynamic and multidimensional (Halliday, 1993). A review of some multidimensional instruments used to measure academic language that are currently available in the research canon are reported in Chapter Two.

Definitions of Key Terms

Text: any passage, spoken or written, of any length, that builds a unified whole.

Register: the way in which language differs according to form (structure), function (purpose), and context in which the language is used.

Discourse: of or pertaining to any unit of communication, spoken or written, that is longer than a sentence.

Academic Language: word, sentence, and discourse level characteristics of language in the academic register

Discourse Types: distinct sub-registers of language that share common structural, functional and contextual factors (i.e., narrative texts versus expository texts)

Lexical: of or pertaining to the individual words of a text, spoken or written.

Grammatical: of or pertaining to the ways in which words, spoken or written, make grammatically correct sentences.

CHAPTER 2: LITERATURE REVIEW

Current academic language assessment methods can be loosely characterized as a) standardized tests, b) structural analyses, or c) mixed approaches. For this review, oral academic language assessments and their respective strengths and limitations are discussed. Drawing from this analysis, an argument is proposed in favor of the development of a user-friendly, multidimensional assessment tool that can accurately and comprehensively measure the spoken academic language of primary grade students.

Assessments of Spoken Academic Language

Assessments of spoken academic language vary widely in the range of techniques they employ, as well as the scope of content areas they target. This section of the review will contain an appraisal of several norm-referenced and criterion-referenced, standardized tests, as well as an evaluation of structural assessment tools for measuring spoken narratives and expositions.

The Standardized Test Approach

An abundance of standardized oral language tests are cited in the literature on children's academic language. Tests are typically administered in pen-and-paper format (i.e., multiple-choice, fill in the blank, short answer) and are occasionally administered orally. Achievement scores on various subtests are calculated according to specific scoring procedures. Those scores are then interpreted as proxy indicators of a student's academic language proficiency level.

Interpretation of test scores varies according to the structure of the test. Whereas norm-referenced tests compare students' achievement to the achievement of other students, criterion-referenced assessments compare student achievement to a predetermined standard. Several norm-referenced and criterion-referenced instruments are described here to illuminate the general structure and mechanics, as well as the strengths and limitations of the standardized test approach to assessing spoken academic language.

Norm-Referenced Tests. An abundance of researchers interested in spoken academic language proficiency have employed tests from the *Woodcock-Johnson* suite of assessments, namely the *Woodcock-Johnson Test of Oral Language (WJ IV-OL)*; Schrank & Wendling, 2018), the *Woodcock Language Proficiency Battery – Revised (WLPB-R)*; Woodcock & Muñoz-Sandoval, 1999), the *Woodcock-Muñoz Language Survey (WMLS)*; Woodcock & Muñoz-Sandoval, 1993). The purpose of the *Woodcock* assessments is to determine and describe an individual's strengths and weaknesses in relation to expressive and receptive language. A series of tests are administered orally in which examiners engage examinees in picture naming, repeating complex instructions, and/or completing sentences to assess for comprehension. Testing typically takes between 15 and 30 minutes. Clusters of these subtests combine to generate composite cognitive academic language proficiency (CALP) scores, which are graded from extremely limited to very advanced based on scores obtained from the norming sample. In this way, CALP scores are interpreted as distal measures of academic language proficiency (Garcia-Bonery, 2011; Hakuta et al., 2000; Laija & Rodriguez, 2006; Sanchez et al., 2013; Tong et al., 2008).

Psychometric properties of the *Woodcock* instruments vary by test edition and form, but reported indices of reliability, internal and external validity are generally adequate. Norming

populations of the *WJ IV-OL* consist of 7,416 people, stratified according to U.S. census data. The *WJ-IV OL* was developed under significant expert review, scored acceptably well in terms of internal consistency reliability (.80-.94) and median cluster reliability (.89-.95), and displayed reasonable patterns of correlation with related cognitive and language tests (Schrack & Wendling, 2018).

The *Kaufman Survey of Early Academic and Language Skills (K-SEALS)*; Kaufman & Kaufman, 1993) is a norm-referenced test designed to evaluate language, articulation, and preacademic concept development of young children. It involves a series of three subtests which independently assess a child's (1) vocabulary, (2) numbers, letters and words, and (3) articulation skills (Cass, 1999). Examiners administer the tests orally with the help of a flip-easel depicting pictures, letters and numbers. Length of testing is between 15 and 25 minutes. Standard scores, percentiles, age equivalents and descriptive categories are presented in the manual. The scores of the first two subtests can be calculated together to provide a composite score that represents the student's Early Academic and Language Skills. The standardization population includes 1,000 children ages 3 through 6 years old, selected through a stratified sampling matrix. Test-retest coefficients were found to be .94, and split-half reliabilities were computed at about .90. for the Early Academic & Language Skills Composite. Researchers and practitioners interpret *KSEALS* scores as indicators of children's early academic and language skills (Uyanik & Kandir, 2014).

The *Test of Narrative Language-2 (TNL-2)*; Gillam & Pearson, 2017) is a norm-referenced, standardized test of narrative comprehension and production which aims to identify children with language disorders. It is intended for use with students ages four through 17. Children listen to three stories with different narrative formats, answer questions about each story, and then either retell the story, or generate a new story. In addition to generating raw

scores for the comprehension questions, examiners score the students' language samples across a number of grammatical criteria (see Table 1 for criterion examples). Age equivalency, percentiles, and standard scores can be calculated for each of the Narrative Comprehension (NC) and Oral Narration (OR) subtests. A composite score can be generated for Narrative Language Ability Index (NLAI).

The reported norming sample for the TNL consists of 1,059 children, stratified by age, gender and race/ethnicity in accordance with U.S. census data. Internal consistency of the items for each subtest fall within an acceptable range ($k = .76$ to $.88$). Test-retest reliability within a two-week testing gap registers between $.80$ and $.90$. Percent agreement scores between $.80$ and $.98$ indicate that interrater agreement is exceptionally good. Measures of sensitivity, specificity, and positive prediction exceed $.85$. In a criterion prediction analysis, correlations with the Spoken Language Quotient (SLQ) of the *Test of Language Development – Primary* (TOLD-P3; Newcomer, & Hammill, 1997) produce coefficients $< .70$, indicating that the TNL is a good measure of general language ability.

Criterion-Referenced Tests. The work of Dr. Paolo Uccelli has been especially influential in recent studies of academic language. In an attempt to expand the field's view of academic language "beyond vocabulary," Uccelli and colleagues (2014) proposed a novel, criterion-referenced instrument to measure core academic language skills (CALS) (Uccelli et al., 2015). In this seminal article, they define CALS to be "the knowledge and deployment of a repertoire of language forms and functions that co-occur with school learning tasks across disciplines" (p. 1). The CALS-1 instrument measures these skills in pre-adolescent learners. Intended for grades four through six, the test examines a set of language skills that facilitate academic text comprehension. Derived from an in-depth literature analysis, CALS-1 items

measure important aspects of academic language proficiency including (1) morphological decomposition, (2) understanding of complex grammar, (3) understanding of school-relevant connectives and discourse markers, (4) anaphoric resolution, (5) argumentative text organization, and (6) academic definitions.

Testing is administered in a 50-minute, paper-and-pencil format to groups of students. An examiner reads words and sentences aloud, asking students to answer various questions. Tasks include multiple-choice, matching, and brief written responses. Items are either dichotomously scored as correct or incorrect, or rescaled to be equally weighted with all other items. Raw scores are then converted to factor scores, and extended scale scores are reported.

Evidence suggests the CALS-1 is a reliable tool ($\alpha = .90$; split-half reliability = .90). An initial study of students ($n = 235$) was conducted by convenience sampling from an urban school in the Northeast U.S. (Uccelli et al., 2014). In an exploratory factor analysis, core academic language task scores loaded onto a single factor, providing evidence of a cohesive underlying construct. CALS-1 scores were found to be predictive of performance on a separate measure of reading comprehension. Additionally, within-grade and between-grade variability was observed in the distribution of students' scores. These findings were replicated in a subsequent study in which English-proficiency designation and SES were found to correlate with between-group variability in CALS-1 scores (Uccelli et al., 2015).

The CUBED assessment is a collection of screening and progress monitoring tools that measure language, decoding, and reading. Although the CUBED adheres to the structural assessment approach defined below, it is described here because the elicitation of language samples is standardized, and the results are compared to grade-level criteria. In other words, the CUBED is a mix between standardized, criterion-referenced and structural assessment

approaches. The Narrative Language Measures (NLM) subtest of the CUBED assessment contains two language comprehension and production measures, the *NLM Listening* and the *NLM Reading* (Petersen & Spencer, 2012, 2016). The *NLM Reading* and the *NLM Listening* were designed to be used in tandem, thereby allowing educators to determine whether a student might benefit from a decoding-oriented intervention, a language-directed intervention, or both. For the retell subtest of the *NLM Reading* and the *NLM Listening*, fictional stories about relatable, primary-age experiences were strategically constructed to contain the structural (e.g., lexical, grammatical and discourse) features representative of typical narrative ability of PreK-3rd grade students. Children retell a grade-level story they either hear read aloud by the examiner, or read independently. Examiners score the retell narratives along multiple dimensions in real time using story-specific scoring rubrics. The scoring rubrics contain an array of items that assess essential lexical, grammatical and discourse features. Raw scores for each item are added together to generate subscale scores, which are the reported metric. CUBED results can be interpreted through a criterion-referenced lens by comparing raw scores with pre-established criterion included in the manual. Students are classified as “at benchmark”, at “moderate risk”, or at “high risk” depending on their performance on each *NLM* subtest.

Psychometric analyses of the *NLM Listening* and the *NLM Reading* indicate acceptable inter-rater (.82 - .95) and alternate forms reliability (.64 - .67). The CUBED manual presents strong evidence of validity, including correlations with related language assessments (e.g., $r = .95$ with the *Renfrew Bus Story*), ability to predict benchmark assessment performance (e.g., $r = .74$ to $.88$ between CUBED language composite and K-3 Measuring Academic Progress (MAP) subscales), and sensitivity to growth.

The Structural Assessment Approach

Linguists and speech-language pathologists frequently conduct structural assessments to measure aspects of language production and comprehension in students suspected of or identified with language impairment (Muñoz et al., 2003). It is important to highlight the ways in which the structural assessment approach contrasts with the language testing approach described in the previous section. Rather than asking students a series of questions and having them respond verbally or in writing, with structural assessments the examiner elicits one or more productive language samples through standardized procedures. Typically, examiners elicit language samples by having students retell a story/passage read to them, or by having students generate a personal story, fictional story, or passage independently. The response is subsequently analyzed for its component features. The language samples are audio recorded, frequently transcribed, and examined at what are commonly referred to as *microstructural* and *macrostructural* levels of analysis.

Microstructural elements are the grammatical and lexical features of a language sample. Microstructural analyses quantify the linguistic features of texts, such as complex noun phrases, adverbs, causal and temporal subordinating conjunctions, coordinating conjunctions, relative clauses, dialogue, length and complexity (Petersen, 2011). Inferences are drawn between different aspects of expressive language and specific metrics. Table 1 provides an overview of some of the more commonly employed metrics and the language features they are frequently paired with in the literature (adapted from Bowles et al., 2020). Researchers analyze these microstructural attributes with the help of computer software programs, such as Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2008).

In contrast, *macrostructural* elements are the discourse-level characteristics of text structure. These language features are expected to vary widely by discourse type (i.e., narrative, expository). Many researchers in this area of study have attempted to draw inferences about macrostructural properties through a parallel, integrative analysis of microstructure measures such as clausal density, productivity, number of T-units and/or MLTU (Nippold & Sun, 2010; Scott & Balthazar, 2010; Price & Jackson, 2015). While microstructural analyses yield valuable information about the grammatical aspects of language, discourse features are equally vital to communication and require investigation in their own right. Lundine (2020) has written plainly that “simply counting the number of language units in a passage may not be a meaningful measure... *more language* may not always be *better*” (p. 157-8). Lundine’s critique and others like it encourage language researchers to explicitly measure text-level, discourse features in their analyses of student language.

In research on narrative discourse, macrostructure is measured directly by quantifying (a) the student’s inclusion of various story grammar elements (Stein & Glenn, 1979), (b) the number of episodes a student produces in the narrative sample (e.g., number of segments containing a problem, a plan/attempt to solve a problem, and a consequence) (McCabe & Peterson, 1984), or (c) some combination of the two. These items are commonly referred to as story grammar complexity, episodic complexity, or in some cases, “narrative quality” (Fey et al., 2004).

There is far less consensus regarding direct assessment approaches to measuring expository macrostructure elements. This is primarily because, in contrast with the uniform, canonical linear form typical of narrative discourse, the form that expository language takes varies widely according to purpose. For example, researchers have identified several distinct

structures that commonly appear in expository contexts, mainly *sequence*, *cause/effect*, *comparison*, *problem/solution*, and *description* (Lundine, 2020).

Table 1

Metrics Commonly Employed in Microstructural Analyses

Language Feature	Indicators
General language productivity	Total # of Utterances (TNU); Total number of words (TNW)
Vocabulary	Number of Different Words (NDW); Number of Different Root Words (NDRW)
Grammatical complexity	Percentage of Utterances Containing Multiple Clauses; Number of T-unit (T-UNIT); Mean length of T-unit (MLT-UNIT); Percentage of grammatical T-unit (GRAM T-UNIT); Number of Clauses; Clause Density (C-DENSITY)
Morphology	Accuracy of Word Inflections
Spelling	Percentage of spelling errors (SPELL)
Writing Conventions	Punctuations

Although there has been a general lack of consensus in the literature over which aspects of expressive language should be included in examinations of child language (Justice et al., 2010), several comprehensive tools have been developed to identify features of linguistic complexity in students' discourse. While the majority of these tools are specifically designed to assess key features of narrative discourse, several attempts have been made to systematically measure expository discourse as well.

It is important to note that structural assessment research has been criticized for frequently bypassing information about the extent to which elicitation and transcription procedures are standardized and implemented with fidelity. Even more concerning is a recurrent absence of information regarding the psychometric soundness of scoring procedures. Due to the automaticity of computerized scoring procedures, reliability and validity indices are generally absent in most research reports (Finestack et al., 2014). Several critical perspectives have highlighted the need for more consistent reporting of elicitation, transcription and coding procedures in child language research (Hadley, 1998; Finestack et al., 2014).

Structural Assessments of Narratives. Petersen and colleagues (2008) developed the *Index of Narrative Complexity (INC)* as a scoring rubric for evaluating microstructural and macrostructural elements of narrative samples. In the spirit of prior criticisms of norm-referenced tests of child language (e.g., Gummarsall & Strong, 1999), they envisioned a tool that would provide useful information about children's narrative development over time and that could inform language intervention efforts. The INC has since been reformulated into the MISL, which stands for *Monitoring Indicators of Scholarly Language* (Gillam et al., 2012). The MISL is a progress monitoring tool that can be used to quantify aspects of students' self-generated narratives. In essence, the MISL is a standardized scoring rubric. It provides minimal direction about procedures for eliciting and transcribing language samples. Audio-recorded language samples are transcribed and analyzed at both microstructural and macrostructural levels using the MISL scoring rubric. The macrostructure subscale of the rubric consists of seven items measuring story elements including character, setting, initiating event, internal response, plan, attempt and consequence. The microstructure subscale includes seven items for measuring literate language structures, namely coordinating and subordinating conjunctions, metacognitive/metalinguistic verbs, adverbs, elaborated noun phrases, grammaticality and tense. Students can earn up to two or three points for each item on the rubric; subscale scores are the reported metric.

Gillam and colleagues (2017) evaluated the psychometric properties of the *MISL* by analyzing the narrative productions of children with language impairments between ages five and eight ($n = 109$). They found inter-scorer reliability for items and subscales ranged from .90 to 1.0. While internal consistency of the microstructure and macrostructure subscales was not adequate, internal consistency for the two subscales combined was sufficient (Cronbach's $\alpha =$

.79) after two of the items from the micro-structure scale were removed (grammaticality and tense). Subsequent research applications of the MISL have provided further evidence of the reliability and validity of the MISL by using a Farsi adaptation of the tool to assess narrative skills of Iranian primary grade children (Beytollahi et al., 2020).

The *NLM Listening* has a companion form called the *NLM Flowchart* (Petersen & Spencer, 2019) designed for narratives generated via alternate elicitation contexts. Much like the *MISL*, the *NLM Flowchart* is a narrative scoring rubric that quickly and efficiently assesses language complexity, narrative structure, and writing conventions. In contrast with the *MISL*, however, standardized elicitation and transcription protocol are included with the *NLM Flowchart*. The instrument has been used in intervention research to observe progressive changes in children's oral (Spencer et al., 2013) and written (Spencer & Petersen, 2018) narrative language. Acceptable scoring agreements (87–96%) and reliability correlations (.57–.69) for the *NLM Flowchart* have been documented. However, an in-depth analysis of the psychometric properties of the tool has not been conducted to date.

Structural Assessments of Expositions. Linguists have documented significant differences in the types of structures contained in expository language, as compared with the language of other registers (Lundine & McCauley, 2016; Schleppegrell, 2001; Snow, 2010). One recent report indicates that register exerts a significant influence on language variables regardless of age, and that text structure, content and domain-specific knowledge moderate this relationship (Hill et al., 2021). For this reason, some language specialists assessing for identification of language impairments have been highly interested in eliciting expository language samples for analysis. Research published to date on this topic does not meet the full criteria for this review; however, two criteria-divergent studies appropriate to the topic should be discussed.

In one study, Westerveld and Moran (2013) investigated differences in linguistic complexity between expository language that is heard and then retold, and expository language that is spontaneously generated. A cross-sectional sampling of primary school ($M = 7.0$ years; $n = 64$); middle school ($M = 11.3$ years; $n = 18$) and high school ($M = 17.6$ years; $n = 18$) students were asked to either talk about their favorite sport, or retell an informational passage read to them about the game of curling. Language samples were transcribed, segmented into T-units, and coded for microstructural elements (T-UNIT, NDW, MLU, CD and PCMZ) using *SALT New Zealand Conventions* (SALT-NZ; Miller et al., 2010). No measures of lexical or discourse-level features were included in this study. The authors reported over 90% agreement in transcription, segmentation, mazing and coding procedures. Grade-level differences were reported for several measures of microstructure complexity, such as clausal density, as well as significant variance in the mean lengths of utterances (MLU's) elicited by retell versus generation conditions.

Lundine and colleagues (2018) developed a scoring rubric to assess spoken summaries of information presented in narrative and expository formats. Fifty adolescents between the ages of 13 and 18 years listened to, and then verbally summarized, one narrative and two expository video-recorded lectures, matched for length and reading level. The expository lectures varied by structure; one was presented in a *compare/contrast* format, and the other a *cause/effect* format. To control for previous knowledge, the subject of all the lectures was a fictitious location called "Lifeland". Each student's summary was transcribed and coded using *SALT* conventions (Miller & Iglesias, 2008). Reliability of transcription and coding procedures was adequate (96% - 100% in a point-to-point comparison). *SALT* software was used to conduct a basic microstructural analysis (MLU and SI). Additionally, the researchers adapted a scoring rubric from a prior study focusing on written language (Westby et al., 2010) to comprehensively assess both

microstructural and macrostructural elements of the verbal summaries. The scoring rubric contained two items for macrostructure and three for microstructure: (1) gist/topic/key sentence/main idea; (2) text structure (e.g., the extent to which the passage is organized and links ideas/main points); (3) content (quantity, accuracy and relevance); (4) conjunctions and signal words to indicate expository subtype; and (5) sentence structure. Students' summaries were scored from 0-4 on each of the five traits. Evidence of psychometric acceptability of the scoring rubric is rudimentary, but promising. The researchers did not report a total inter-scorer agreement value in their report; however, they did report that 95.6% of derived scores matched or differed by only 1 point, and that perfect agreement was achieved on 52% of scores.

Strengths and Limitations of Current Approaches

Standardized Tests

Standardized tests of oral academic language appear to display good evidence of reliability and validity. Norm-referenced tests like the *WJ-TOL*, the *K-SEALS*, and the *TNL-2* can effectively distinguish between children who have significantly more difficulties with academic language than their peers, making them suitable instruments for diagnostic purposes. Evidence regarding the reliability and validity of tests of *written* academic language, however, are more variable. Internal consistency and concurrent validity indices of norm-referenced writing tests such as the *SAT-9* and the *TOWL-4* are adequate, albeit weaker than their oral language test counterparts. But comparatively weak interrater reliability estimates indicate that scorer bias may interfere with objective measurement of student writing when using these types of assessment tools.

One shortcoming associated with norm-referenced tests is that the information these tests provide is extremely limited in application to intervention and instruction. Researchers have noted that little use has been made of tests like the *TNL-2* in clinical or educational settings because they are not directly linked with any validated intervention packages or strategies (Hayward et al., 2008b). This shortcoming may be especially problematic in terms of intervening to enhance academic language proficiency, since norm-referenced tests tend to assess global skills repertoires (i.e., oral language *in general* rather than academic language *in particular*).

Moreover, standardized, norm-referenced tests tend to require considerable time and resources to administer. One major advantage of criterion-referenced tests over norm-referenced tests is that criterion-referenced tests are typically more flexible and less resource-intensive. For example, a single examiner can administer the *CALS-I* to large groups of students at a time, making it an extremely helpful tool for screening purposes. The *NLM-Listening* and the *NLM-Reading* take less than 5 minutes each to administer, and with multiple retell stories available can be used to progress monitor growth over time.

An additional strength of criterion-referenced tests is that they tend to be narrower in scope than norm-referenced tests. By focusing exclusively on a targeted constellation of features, researchers can better identify and understand the phenomenon of interest, ultimately leading to better-informed, more effective instruction. This is true of the *CALS-I* and the *CUBED* assessments, which were developed in accordance with highly specific operational definitions of the academic language register. The *NLM-Listening* and the *NLM-Reading* are even more fine-tuned to a specific construct, since they pointedly assess a single discourse type (e.g., narrative academic language) across multiple domains of linguistic elements.

Structural Assessments

There are considerable benefits associated with the structural assessment approach to measuring academic language. The scope of data that can be generated through this method is limited only by opportunity and access. It typically only takes between five and ten minutes to elicit a language sample, making it an ideal measurement approach for progress monitoring. In the absence of explicit, standardized testing procedures, minimal training and materials are required to gather data for analysis. However, there is a major drawback associated with that convenience. A lack of standardized procedures for eliciting student language samples may contribute to inconsistent data, and thereby unreliable results. As previously discussed, too many structural assessments published to date do not report the psychometric properties of their elicitation procedures. This shortfall has rightly been called into question by language researchers (Finestack et al., 2014). The validation of this assessment approach depends largely on the extent to which research methods can be carried out with rigorous standardization procedures.

As previously mentioned, structural analyses tend to emphasize grammatical aspects of language, often at the expense of more explicit considerations of lexical and discourse-level features. Whereas reliable, validated methods of assessing narrative macrostructure elements in spoken language are well established in the research literature (e.g., story grammar elements), explicit measures of expository macrostructure are still being tested and developed. While there are some promising strategies being used to assess expository structures, there is no assessment tool available that employs these strategies in a format that is accessible for generalized use. Hence, there is still significant work to be accomplished in this area.

Two significant drawbacks associated with the structural assessment approach are also evident. First, they require a considerable amount of training and resources to transcribe, score and interpret language samples into meaningful information. Microstructural analyses are typically conducted by individuals with extensive knowledge of linguistic structures and functions, and/or with the help of computer software programs. Macrostructural analyses may require even more interpretive training, since discourse features tend to be more abstract and difficult to capture than word-level and sentence-level features. This is especially true of expositions, which take on a variety of structural forms according to their purpose.

Second, in conducting structural assessments of student language, linguists discriminate between microstructural and macrostructural elements to capture meaningful information about the many dimensions that affect overall quality. While these categories are helpful, it appears that researchers have paid considerably more attention to the analysis of microstructural elements at the expense of developing a systematic, comprehensive approach to analyzing both the micro- and macrostructural elements of texts. This is especially true regarding expository discourse. While there are some promising areas of exploration, much is left to be discovered. There is a vital need for more precise measures of academic language that can simultaneously capture multiple dimensions of the construct. While current approaches to measuring academic language are valuable for a wide variety of purposes, what remains to be developed is a suite of measures that are *comprehensive* in scope, discourse-specific, easy to interpret, and informative for instructional decision-making.

Purpose of the Current Research Study

The purpose of this research is to examine the psychometric properties of two discourse analysis tools designed to measure children’s academic language—the *Narrative Language Measures (NLM) Flowchart* and the *Expository Language Measures (ELM) Flowchart*. The research questions of the current study are:

1. To what extent, if any, do kindergarten through 3rd grade students’ oral academic language skills, as measured by the *NLM Flowchart* and *ELM Flowchart*, vary by students’ grade level?
2. When two scorers independently use the Flowcharts, what is the interrater reliability?
What is the level of agreement among scorers?
3. What is the factor structure of the Flowcharts?
4. What is the internal consistency reliability of the identified factors?
5. To what extent do the factor scores relate to scores derived from a norm-referenced test of academic language?

According to the *Standards for Psychological and Educational Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), “statements about validity should refer to particular interpretations for specified uses” (p. 11). The discourse analysis tools presented in this report were developed to capture meaningful information about the academic features of language produced by young students. Accurate, useful data will help researchers better understand the academic language construct and further inform the development of targeted interventions to address students’ specific language needs. Thus, assertions about the validity of these instruments are made in direct relation to that end.

Importance of the Current Research Study

Given their importance in the reading and writing development of school age children, academic language skills are worthy of early, intensive instruction. However, until the educational community has a better understanding of exactly what, when and how academic language develops, it will be challenging to maximize instructional efforts. The aim of this study, therefore, is to examine a suite of innovative assessment tools designed to advance the research on the academic language of primary school students. Specifically, psychometric properties are being investigated to determine the extent to which these instruments can generate accurate, useful data, and thereby inform the development of empirically-based academic language interventions and instructional strategies. It is assumed that additional research and development would be necessary to refine and adapt these tools to evaluate the effectiveness of instruction and/or quantify students' rates of improvement or responsiveness to instruction. Such work is outside the scope of this study.

CHAPTER 3: METHOD AND PROCEDURES

The current research study is an investigation of the psychometric properties of two academic language discourse analysis tools – the *Narrative Language Measure (NLM) Flowchart* and the *Expository Language Measure (ELM) Flowchart* (see Appendices). The measures were designed to be direct (e.g., requiring minimum inference), comprehensive, and appropriate for use in educational settings. The *NLM Flowchart* and *ELM Flowchart* can be used to measure spoken or written academic language; however, this study focuses exclusively on spoken academic language to prioritize application to early (e.g., pre-orthographic) intervention efforts.

We define academic language as a collection of distinct lexical, grammatical and discourse features that are frequently encountered and employed in school settings, and do not occur at high rates in the conversational language of primary age students. The idea map in Figure 1 illustrates the framework in which the *NLM* and *ELM Flowcharts* were conceptualized.

Academic language can be differentiated into narrative and expository (informational) discourse types. All three levels of academic language (lexical, grammatical and discourse) within each discourse type can be analyzed. Lexical and grammatical features are measured via items in the *Language Complexity* subscales. Items in the *Narrative Structure* and the *Passage Structure* subscales reflect narrative and expository discourse features, respectively.

Prior to the current study, a version of the *NLM Flowchart* had been used to measure student-generated personal stories (Spencer et al., 2015) and written stories (Kirby et al., 2021;

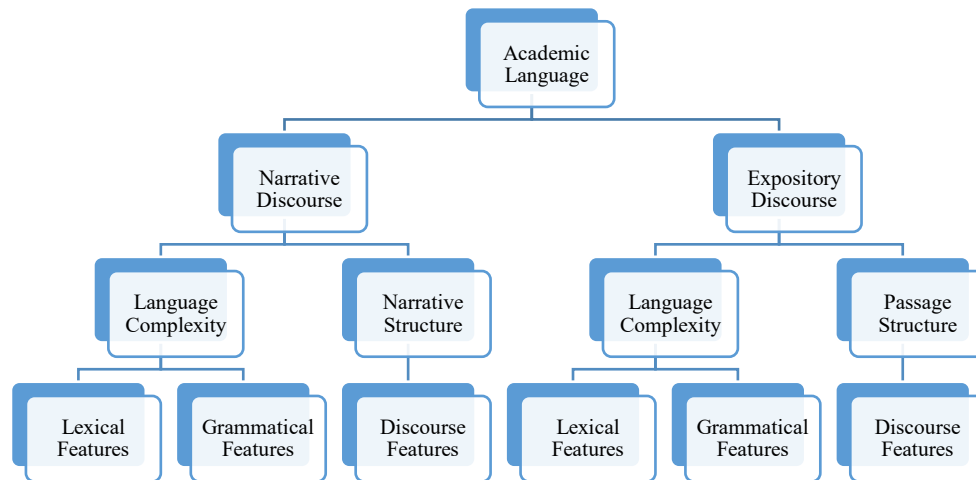


Figure 1
Conceptual Map of Academic Language via the NLM and ELM Flowcharts

Spencer & Petersen, 2018). As a companion to the *NLM Listening and Reading*, the *NLM Flowchart* includes similar items related to story structure (i.e., character, setting, problem, emotion, plan, attempt, consequence, ending) and language complexity (i.e., causal and temporal ties), but was recently enhanced to include complex grammatical and lexical features that are too difficult to capture in real time like the *NLM Listening and Reading* demand.

In contrast to the *NLM Flowchart*, the *ELM Flowchart* had no predecessor; however, the general layout of the tool is similar to the *NLM Flowchart*. Initial items for the *ELM Flowchart* were generated through an exhaustive review of literature related to expository academic language. Indispensable to this review were corpus analyses of the language structures encountered in academic textbooks and in the productive academic language (oral and written) of children and adolescents (Biber et al., 1999). Ultimately, most of the items from the *Language Complexity* subscale of the *NLM Flowchart* were retained, but the *Passage Structure* subscale contained novel items relevant to the expository discourse research literature.

Language Complexity subscale	Narrative Structure subscale (NLM only)	Passage Structure subscale (ELM only)
<ul style="list-style-type: none"> • Relative Pronouns • Verb/Noun Modifiers • Vocabulary/Rhetoric • Temporal Ties • Causal Ties • Dialogue (<i>NLM Flowchart</i>) or Transition Words (<i>ELM Flowchart</i>). 	<ul style="list-style-type: none"> • Episode Complexity • Character • Setting • Problem • Plan/Attempt • Consequence • Ending • Sequence • Emotion 	<ul style="list-style-type: none"> • Main Idea • Information Units • Definitions and Examples • Passage Cohesion • Concluding Statement • Exposition Type

Figure 2
Subscales and Items of the NLM and ELM Flowcharts

The *NLM* and *ELM Flowcharts* were iteratively developed and refined over the course of two years prior to this study. In multiple cycles, at least two raters used the draft versions of the Flowcharts to score sets of 50-100 language samples. Based on interrater reliability and item total correlation results, items were either eliminated or revised. Additionally, this iterative refinement informed the development of the *NLM Flowchart* and *ELM Flowchart* scoring guides that were used to score the language samples in this study. The final versions of the Flowcharts contain the subscales and items presented in Figure 2.

Participants

Participants were recruited from 60 before/after school care and summer care programs operated by the school district or the parks and recreation department. The particular county was strategically chosen because its student population roughly represents the national student population. Any student currently in or entering (if data were collected in summer) into K-3rd grade in the subsequent fall semester was eligible to participate. Enrollment of sites and students was rolling and took place over 15 months. To enroll participants and conduct informed consent,

research assistants (RAs) visited each site during pick up times to speak to caregivers about the study. Spanish speaking RAs were available to speak to Spanish-speaking caregivers as needed. While speaking to caregivers, RAs explained the study in detail and requested permission for their child to participate. At the time informed consent was collected, caregivers completed a brief demographic survey to ascertain children's race/ethnicity, languages, and special education status. Demographic questionnaires were provided to caregivers in English and Spanish. Data from the demographic questionnaires was then entered into an online data repository. The research team de-identified participants by utilizing their county-administered student ID numbers instead of first/last names on all recordings and study materials. This ensured confidentiality for student data. Risks associated with participation in the study were minimal, except for potential loss of confidentiality through voice recordings. In total, 1,179 K-3rd grade students participated in the study, but demographic questionnaires were completed for only 1,040 students. See Table 2 for a summary of demographic characteristics.

Research Team

A total of 11 RAs collected language samples over the course of the study. Staff consisted of four full-time RAs with undergraduate degrees and 7 part-time, undergraduate RAs. Prior to data collection, RAs attended an initial 2-hour training with an experienced *Woodcock-Johnson* test administrator, as well as a 2-hour training on language elicitation procedures. A check out procedure was employed to ensure that trainees correctly followed assessment and language elicitation protocols. In addition, regular fidelity checks occurred to ensure the maintenance of study procedures.

A group of *NLM Flowchart* and *ELM Flowchart* scorers was assembled, which consisted of six staffed researchers with undergraduate degrees and two undergraduate research assistants. Scorers attended an initial 2- to 3-hour workshop provided by the author of this paper in which each measure was introduced individually, with examples and non-examples. Scorers were then assigned a battery of practice examples to complete individually. Following completion of the practice samples, an additional 1-hour follow-up session was scheduled with the researcher or a previously trained scorer to discuss the trainees' scores and clarify scoring procedures. Newly trained scorers met with more experienced scorers as needed. Ongoing calibration meetings were conducted every 1-2 weeks, in which all scorers met to discuss difficult samples and developed guidelines/criteria for making fine-grained decisions about specific items.

Flowchart Materials and Standardized Procedures

Data for the variables of interest were obtained in five phases. The order in which the activities were carried out is shown in Table 3.

Data Collection

For Phase 1, RAs administered subtests one (Picture Vocabulary), two (Oral Comprehension) and three (Understanding Directions) of the *WJ-IV*, which together form a composite.

Administration was conducted individually at pre-determined locations in the school/center, and took approximately 20 minutes to complete. RAs recorded students' responses during the assessments and calculated total raw scores later. Raw scores, standard scores and CALP scores from the *WJ-IV* are used as criterion measures of academic language, compared to the results of the Flowcharts.

Table 2*Demographic Characteristics of Participants (n = 1,040)*

Demographic	N (percentage)	
<i>Gender</i>		
Female	525	(50.5%)
Male	515	(49.5%)
<i>Grade</i>		
K	282	(27.1%)
1	257	(24.7%)
2	279	(26.8%)
3	222	(21.3%)
<i>Age</i>		
5.0 - 5.9	134	(12.9%)
6.0 - 6.9	282	(27.1%)
7.0 - 7.9	267	(25.7%)
8.0 - 8.9	251	(24.1%)
9.0 - 9.9	103	(9.9%)
10.0 - 10.9	3	(0.3%)
<i>Ethnicity</i>		
White	396	(38.1%)
Hispanic/Latino	371	(35.7%)
African American	354	(34.0%)
Asian American	42	(4.0%)
Native American	5	(0.5%)
Other	33	(3.2%)
<i>Language status</i>		
Language spoken at home		
English only	831	(79.9%)
Spanish only	150	(14.4%)
English and Spanish (bi-lingual)	42	(4.0%)
Other	17	(1.6%)
Language most comfortable		
English only	958	(92.1%)
Spanish only	37	(3.6%)
English and Spanish (bi-lingual)	42	(4.0%)
Other	3	(0.3%)
Reported language concerns		
Yes	161	(15.5%)
No	879	(84.5%)
<i>Special Education Status</i>		
Reported Individualized Education Plan (IEP)	140	(13.5%)
<i>Mother's Highest Education</i>		
Elementary	17	(1.6%)
Some high school, no diploma	30	(2.9%)
High school education	162	(15.6%)
Some college, no degree	341	(32.8%)
Bachelor's degree	286	(27.5%)
Master's degree	181	(17.4%)
Doctoral degree	23	(2.2%)
<i>Father's Highest Education</i>		
Elementary	31	(3.0%)
Some high school, no diploma	70	(6.7%)
High school education	305	(29.3%)
Some college, no degree	292	(28.1%)
Bachelor's degree	229	(22.0%)
Master's degree	97	(9.3%)
Doctoral degree	16	(1.5%)
Total Sample	1040	

Table 3*Five Phases of Data Collection and Scoring*

Phase	Activity
1	Woodcock-Johnson IV (WJ-IV) – Test 1, 2 & 6
2	Elicit 2 Expository Retell Oral (ERO) + 2 Expository Generation Oral (EGO)
3	Elicit 2 Narrative Retell Oral (NRO) + 2 narrative Generation Oral (NGO)
4	Language sample transcription
5	Language sample scoring with the <i>NLM</i> and <i>ELM Flowcharts</i>

For Phases 2-4, researchers administered standardized procedures for eliciting high-quality, academic language samples (see Appendix E for an elicitation script example). Each student provided language samples across three 10- to 15-minute sessions. We used a spaced procedure to avoid any potential priming effects that might occur, thereby influencing the types of responses students might provide. Students were only permitted to participate in one session per day. In a series of two sessions, each student had the opportunity to produce two retell expository oral language samples and two generated expository language samples (Phase 2), and two retell narrative oral language samples and two generated narrative oral language samples (Phase 3).

At the beginning of each session, the RA showed the student three randomly selected sets of narrative photos or expository photos. Students were asked to select the set they wanted to talk about. The examiner first read aloud a story or informational passage corresponding with the chosen photo set, which the student was then asked to retell. This procedure was repeated with a second elected photo set. After the third set of photos was selected, examiners asked students to generate their own story or information about the pictures they selected, which was repeated with a fourth set of photos. All elicitations were recorded on audio devices and the written samples were collected.

Transcription and Coding

In Phases 4 and 5, the recorded language samples were transcribed in accordance with corpus linguistic standards, and then scored using the *NLM* and *ELM Flowcharts*. Fidelity checks were conducted regularly to ensure elicitation and transcription integrity. An independent RA listened to 26% ($n = 1,060$) of recordings of language sample elicitations and used a checklist to document adherence to the protocol. Using this procedure, elicitation fidelity was determined to be 99%. Additionally, 24% ($n = 955$) of the total samples were transcribed by a second, independent RA. A third person then reviewed the first and second transcriptions, calculated percent agreement between the two, and documented adherence to transcription procedures for each transcriber. A mean transcription fidelity score for transcriber one was calculated at 99%. Transcribed content was identical between transcribers an average of 93% of the time. Any language samples with an agreement score below 80% underwent a reconciliation process, wherein a third transcriber listened to recordings and used best judgement to decide on a final transcription.

Overview of Data Analysis Strategy

Estimates of Validity

A CFA was conducted to test two-factor measurement models of the *NLM* and *ELM Flowcharts* for spoken academic language. Models of the proposed factor structures are displayed in Figure 3. The *exposition type* and *information units* items of the *ELM Flowchart* were not included in the factor analysis for the following reasons. First, *exposition type* is a categorical variable. CFAs are based on a variance-covariance matrix, which assumes at least ordinal variables. Second, the range of values for the *information units* item (0 – 66) is much larger than any of the other indicators of the *ELM Flowchart* (i.e., 0 – 4). When dealing with

items on different scales, standardized factor scores can be used to estimate the factor structure on a calculated scale. However, given that expository discourse is, by definition, an ordered assemblage of superordinate and subordinate information units (Mosenthal, 1985), it was anticipated that the *information units* item would likely have a factor loading value far greater than 1. Therefore, the researchers chose to restrict the focus of the current study to the structure and reliability of the other *Passage Structure* components.

Finally, total score correlations between *NLM* and *ELM Flowchart* scores and *WJ-IV OL CALP* scores were calculated to indicate what type of relationship exists between the *Flowcharts* and a norm-referenced, standardized assessment.

Estimates of Reliability

Two researchers independently scored 25% of the language samples using the *Flowcharts*. Point-by-point percent agreement scores and Cohen's kappa coefficients (Cohen, 1960) were calculated to indicate the level of agreement among scorers. To determine the internal consistency reliability of the factors identified through the CFA, Cronbach's alpha coefficients for each factor were also calculated.

Interpreting Psychometric Properties of Productive Language Assessments

It should be emphasized here that academic language is a multidimensional, rather than unidimensional, construct. According to Law and colleagues (1998), a multidimensional construct "consists of a number of interrelated attributes or dimensions, and exists in multidimensional domains. In contrast to a set of interrelated unidimensional constructs, the dimensions of a multidimensional construct can be conceptualized under an overall abstraction,

and it is theoretically meaningful and parsimonious to use this overall abstraction as a representation of the dimensions” (p. 741).

The development of academic language skills depends on an integration of interrelated lexical, grammatical, and discursive abilities (Schleppegrell, 2001). These prerequisites are not necessarily causally related, and should not be expected to correlate strongly. In accordance with this observation about the structure of academic language, it is not anticipated that there will be especially high inter-correlations between factor indicators. The relationship between the latent variables and the indicators can be characterized by what some researchers refer to as a formative, rather than reflective, affiliation (Bollen, 2011). To interpret the psychometric properties of the *Flowcharts* in the conclusion section of this study, this conceptual distinction will be applied and explored.

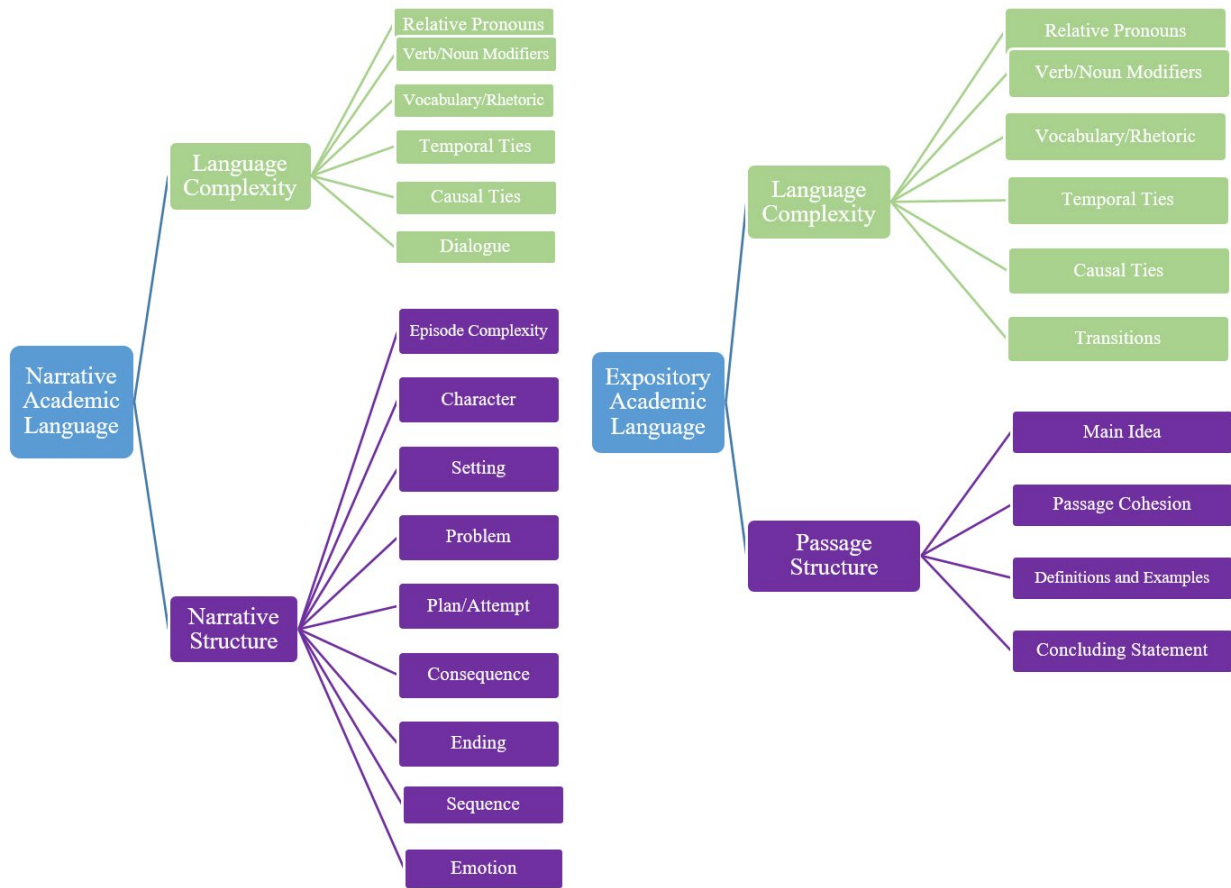


Figure 3
Proposed Factor Structure of Narrative and Expository Academic Language

CHAPTER 4: FINDINGS AND RESULTS

Descriptive Statistics

Two factors were specified for each instrument: *Language Complexity* (Factor 1) and *Narrative Structure* (Factor 2) for the *NLM Flowchart*, and *Language Complexity* (Factor 1) and *Passage Structure* (Factor 2) for the *ELM Flowchart*. Scores were highest for items that loaded on the *Narrative Structure* factor (generation samples $M = 15.10$, $SD = 1.10$; retell samples $M = 18.02$, $SD = 1.81$; $Range = 0 - 34$) and lowest for items on the *Passage Structure* factor (generation samples $M = 2.15$, $SD = .58$; retell samples $M = 1.99$, $SD = .59$, $Range = 0 - 10$). For narrative generation samples, individual items loading onto the *Narrative Structure* factor had mean scores ranging from .78 ($SD = 1.04$, $Range = 0 - 3$) for the *emotion* item to 3.10 ($SD = 1.96$, $Range = 0 - 8$) for *episode complexity*. These same two items defined the lower and upper limits for retell samples at .73 ($SD = 1.03$, $Range = 0 - 3$) for *emotion* and 3.92 ($SD = 2.18$) for *episode complexity*. On the *Passage Structure* factor, indicator mean scores ranged from .04 ($SD = .20$) for *concluding statement* to 1.42 ($SD = .69$) for *passage cohesion*. For *Passage Structure* items, little difference was observed between mean score values across task types.

Overall, mean scores for the *Language Complexity* factor from each instrument differed based on task type and genre. Mean scores were slightly higher for narrative retell samples ($M = 3.75$, $SD = .85$) than for narrative generation samples ($M = 2.29$, $SD = .70$). In contrast, mean scores for expository generation samples ($M = 3.98$, $SD = .90$) slightly surpassed expository retell samples ($M = 3.53$, $SD = .83$). Individual item means from the *Language Complexity*

subscale of the *NLM Flowchart* for generation samples ranged from .14 for the items *relative pronouns* ($SD = .45$) and *dialogue* ($SD = .43$), to .85 for the item *verb/noun modifiers* ($SD = 1.16$). Retell sample mean scores ranged from .17 ($SD = .47$) for the *relative pronouns* item, to 1.35 ($SD = 1.29$) for the *verb/noun modifiers* item. Interestingly, the *dialogue* item mean score differed between generation and retell samples with a score of .44 ($SD = .70$) for retells and .14 (.43) for generations (mean difference score = .30). For expository generation samples, individual item mean scores ranged from .02 ($SD = .21$, Range 0 – 4) for the *transitions* item to 1.26 ($SD = 1.32$, Range = 0 – 3) for *verb/noun modifiers*. These same two items defined the upper and lower limits for retell samples at .01 ($SD = .13$, Range = 0 – 4) for *transitions* and 1.06 ($SD = 1.26$) for *verb/noun modifiers*. Hence, the *Language Complexity* item with the highest scores across genre and task type was the *verb/noun modifiers* item. Of all the items loading onto the *Language Complexity* factor, the lowest scores came from the *dialogue* and *transitions* items of the *NLM Flowchart* and the *ELM Flowchart*, respectively.

Skewness and kurtosis values indicated normal distributions for four of twenty indicators: *episode complexity*, *setting*, *emotion*, and *passage cohesion*. Eleven indicators had skewed distributions compared to a normal distribution. Skewness values for *problem* (-1.84, -1.79), and *consequence* (-1.05, -1.14) were left-skewed. Distributions were right-skewed for nine indicators, namely *relative pronouns* (3.61, 3.26, 2.01, 2.45), *vocabulary/rhetoric* (2.82, 1.21, 1.08, -0.84), *temporal ties* (1.79, 1.45, 1.46, 2.27), *causal ties* (1.74, 1.55, 0.78, 1.53), *dialogue* (3.35, 1.30), *transitions* (12.65, 16.99), *main idea* (2.51, 2.93), *examples and definitions* (1.46, 1.54), and *concluding statement* (4.64, 4.99). Kurtosis values indicated non-normal distributions for eleven items. Platykurtic distributions were identified for three items, namely *verb/noun modifiers* (-1.02, -1.71, -1.73, -1.51), *sequence* (-1.10, -0.37), and *ending* (-1.74, -1.93). Eight

indicators displayed evidence of leptokurtic distributions, including *relative pronouns* (14.24, 11.65, 3.22, 5.77), *vocabulary/rhetoric* (8.94, 0.56, 0.14, -0.22), *temporal ties* (2.21, 1.09, 0.74, 3.99), *causal ties* (2.15, 1.44, -1.0, 1.04), *dialogue* (10.67, 0.34), *transitions* (184.63, 307.81), *concluding statement* (19.58, 22.91), and *main idea* (5.26, 7.70). See Table 4 for individual items means, standard deviations, skewness and kurtosis values.

Differences across Grade Levels

NLM Flowchart mean scores were observed to increase across grade levels. The average *NLM Flowchart* scores was 14.2 for Kindergarteners, 18.4 for 1st graders, 22.4 for 2nd graders, and 23.9 for 3rd graders. An increasing trend for mean *ELM Flowchart* scores was also observed: 7.0 for Kindergarteners, 9.1 for 1st graders, 11.0 for 2nd, and 11.2 for 3rd graders. Figure 4 depicts the changes across grades in bar graph form.

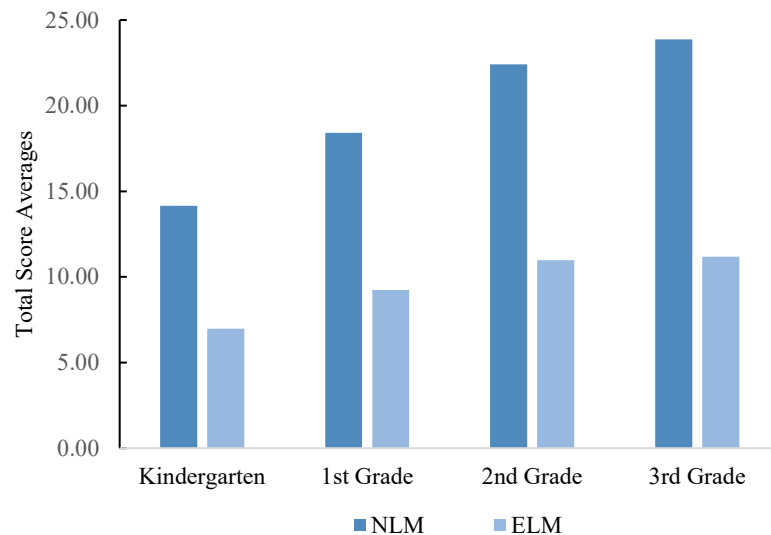


Figure 4
NLM and ELM Scores across Grade Levels

Table 4

Descriptive Statistics

Instrument	Factor	Indicator	Range	Generation			Retell		
				<i>M(SD)</i>	<i>Skewness</i>	<i>Kurtosis</i>	<i>M(SD)</i>	<i>Skewness</i>	<i>Kurtosis</i>
NLM Flowchart	Language Complexity	Relative Pronouns	0 - 3	.14 (.45)	3.61	14.24	.17 (.47)	3.26	11.65
		Verb/ Noun Modifiers	0 - 3	.85 (1.16)	0.81	-1.02	1.35 (1.29)	0.09	-1.71
		Vocabulary/ Rhetoric	0 - 3	.21 (.50)	2.82	8.94	.67 (.88)	1.21	0.56
		Temporal Ties	0 - 3	.48 (.86)	1.79	2.21	.58 (.89)	1.45	1.09
		Casual Ties	0 - 3	.47 (.82)	1.74	2.15	.54 (.86)	1.55	1.44
		Dialogue	0 - 3	.14 (.43)	3.35	10.67	.44 (.70)	1.30	0.34
		<i>Language Complexity</i>	<i>0 - 18</i>	<i>2.29 (.70)</i>			<i>3.75 (.85)</i>		
	Narrative Structure	Episode Complexity	0 - 8	3.10 (1.96)	-0.27	-0.85	3.92 (2.18)	-0.50	-0.67
		Character	0 - 3	.97 (.61)	0.26	0.55	1.77 (1.38)	-0.39	-1.72
		Setting	0 - 3	.81 (.82)	0.68	-0.37	1.03 (.91)	0.41	-0.80
		Problem	0 - 4	2.62 (1.02)	-1.84	2.14	2.71 (1.04)	-1.79	2.39
		Sequence	0 - 3	1.43 (.91)	-0.93	-1.10	1.54 (.85)	-1.26	-0.37
		Plan/ Attempt	0 - 4	2.31 (1.31)	-0.98	-0.58	2.80 (1.19)	-1.41	1.13
		Consequence	0 - 4	2.27 (1.25)	-1.05	-0.47	2.53 (1.27)	-1.14	0.05
		Ending	0 - 2	.81 (.93)	0.38	-1.74	.99 (.97)	0.02	-1.93
		Emotion	0 - 3	.78 (1.04)	0.82	-0.90	.73 (1.03)	0.91	-0.79
		<i>Narrative Structure</i>	<i>0 - 34</i>	<i>15.10 (1.10)</i>			<i>18.02 (1.81)</i>		
		Total	0 - 42	17.38 (1.80)*			21.76 (2.66)**		
ELM Flowchart	Language Complexity	Relative Pronouns	0 - 3	.41 (.80)	2.01	3.22	.30 (.67)	2.45	5.77
		Verb/ Noun Modifiers	0 - 3	1.26 (1.32)	0.24	-1.73	1.06 (1.26)	0.50	-1.51
		Vocabulary	0 - 3	.73 (.92)	1.08	0.14	.84 (.91)	0.84	-0.22
		Temporal Ties	0 - 3	.60 (.99)	1.46	0.74	.37 (.83)	2.27	3.99
		Casual Ties	0 - 3	.96 (1.18)	0.78	-1.00	.96 (1.18)	1.53	1.04
		Transitions	0 - 4	.02 (.21)	12.65	184.62	.01 (.13)	16.99	307.81
		<i>Language Complexity</i>		<i>3.98 (.90)</i>			<i>3.53 (.83)</i>		
	Passage Structure	Main Idea	0 - 3	.22 (.56)	2.51	5.26	.18 (.53)	2.93	7.70
		Definitions & Examples	0 - 3	.47 (.88)	1.46	0.42	.46 (.89)	1.54	0.72
		Passage Cohesion	0 - 3	1.42 (.69)	0.15	-0.17	1.30 (.76)	0.09	-0.37
		Concluding Statement	0 - 1	.04 (.20)	4.64	19.58	0.04 (.19)	4.99	22.91
		<i>Passage Structure</i>		<i>2.15 (.58)</i>			<i>1.99 (.59)</i>		
		Total		6.13 (1.48)***			5.52 (1.42)****		

* $n = 1,966$; ** $n = 1,966$; *** $n = 2,008$; **** $n = 1,947$

Factor Structure of NLM and ELM Flowcharts

A CFA was conducted to test two-factor measurement models of the *NLM* and *ELM Flowcharts*. Items were initially specified based on an evaluation of the extent to which they theoretically fit into the latent constructs. A two-factor model was hypothesized to reflect the distinctions between microstructure (grammatical/lexical features) and macrostructure (discourse features) typically employed by researchers and practitioners in the communication sciences. Investigations involved modeling item scores using Mplus software (Version 8; Muthén & Muthén, 2017). All models were estimated using the mean-and-variance-adjusted weighted least squares (WLSMV) estimator. This estimator was chosen because it is suitable for employment with data that do not display multivariate normality. Samples (Level 1) were nested in school sites (Level 2), with a mean of 207 samples per school site ($SD = 111$). Nesting effects were controlled using the Mplus feature of TYPE _ COMPLEX to methodologically minimize the influence of each individual school site and to maximize the study's generalizability.

The CFA was assessed for exact fit via a maximum likelihood (ML) χ^2 appraisal. Exact model fit would be concluded if a non-significant χ^2 value ($p > .05$) was found. In case of model misspecification, approximate fit was evaluated using standardized root-mean-square residual (SRMR), comparative fit index (CFI)/ Tucker-Lewis Index (1973, TLI), and root-mean-square error of approximation (RMSEA). Hu and Bentler (1999) recommend using these calculations with ML methods to determine the extent to which a model displays sufficient evidence of fit for model misspecification. According to these studies, approximate fit may be assumed if a model achieves the following fit index values: a SRMR $< .08$ (primary criterion) and either a CFI/TLI $> .95$ or an RMSEA $> .06$ (secondary criteria). Additionally, Brown (2015) suggested factor loadings of individual items must be greater than or equal to .30 or .40 to be

considered acceptable in applied research. We applied these criteria to guide decision-making regarding adjustments (e.g., item deletions or correlations) to the proposed model.

To determine model specifications, researchers reviewed CFA model fit indices, individual item factor loadings, and modification indices for the original models (Model 1 of narrative and expository academic language, respectively) in Mplus (Version 8; Muthén & Muthén, 2017). Item decisions were made by first looking at modification indices to identify indicators that were either strongly correlated with another item, or that were potentially crossloading onto the non-indicated factor. Modification index values were interpreted in light of our conceptual understanding of the indicators based on research, how they might relate to the overall construct, and how they might interact with each other in productive language. Large modification index values that were consistent with theory and prior research were tested through modified models. Modification index values that were significant but inconsistent with prior research were not tested (i.e., items were not removed or evaluated on another factor). Researchers then reviewed factor loadings to determine whether there were any items that, if removed, might make the instrument more accurate in capturing the identified construct.

Evidence of bi-dimensionality was compared against uni-dimensionality, but additional factors were not explored in the course of this study. As previously discussed, there is an abundance of theory and research in the speech/language cannon that describes academic language in terms of microstructural (lexical/grammatical) and macrostructural (discursive) features, lending support to a bi-dimensional structure (Petersen, 2011). Moreover, internal consistency is affected by the number of items in a scale in that the less items the scale contains, the greater the correlations need to be between items in order for alpha values to be significant (Brown, 2015). Since each *Flowchart* subscale contains a relatively small number of items to

start with, it was thought that further dividing the items into additional factors would likely have a detrimental effect on the overall fit of the model.

For expository academic language conceived through the two-factor *ELM Flowchart* model, four additional models were specified. For the two-factor *NLM Flowchart* model, analyses were conducted for two additional models. Table 5 displays fit index values for the different models.

ELM Flowchart Model Specifications

Modification indices for *ELM Flowchart* Model 1 signaled notable correlations between the *main idea* and *passage cohesion* indicators (generation M.I. = 24.86; retell M.I. = 22.29). These two indicators are conceptually interdependent; *passage cohesion* assesses the extent to which the information units support an explicitly stated main idea. Hence in Model 2, researchers controlled for this correlation. Upon further review, it was hypothesized that *passage cohesion* may actually capture *main idea* entirely within its operational definition.

Researchers reviewed factor loadings and found that the data were consistent with this hypothesis: *main idea* loaded onto the specified factor with less power than *passage cohesion* (difference score for retell = .23; difference score for generation = .13). Hence in Model 3, researchers tried deleting the *main idea* indicator to assess any differential changes in fit index values. Modification index values also suggested that the *vocabulary* item crossloaded across Factors 1 and 2 (generation M.I. = 55.84; retell M.I. = 67.74). This relationship was somewhat expected. Vocabulary is closely related to information units in that the more an individual knows about a specific topic, the greater their vocabulary knowledge. Conceptually, this item could be grouped with the *Passage Structure* or *Linguistic Complexity* factor; however, indicators that cross-load onto multiple factors can pose threats to discriminant validity. Therefore in Model 4,

vocabulary was grouped within the *Passage Structure* factor to assess differences of fit. Finally, in Model 5 the *transitions* item was removed due to poor loading onto the indicated factor (generation = 0.07; retell = 0.18).

There were several unexpectedly high modification index values that were inconsistent with prior research. *Causal ties* and *temporal ties* showed some evidence of correlation for retell samples only (M.I. = 31.62). Students may have used these structures in tandem when retelling expository passages because they were modeled together in the retell passage. Correlations between *concluding statement* and *main idea* scores were noted (retell M.I. = 28.97; generation M.I. = 51.88). However, the *concluding statement* item loaded poorly onto the identified factor (retell = 0.27; generation = 0.28), so it was hypothesized that altering the model to account for this correlation would not significantly impact the overall model fit. Finally, the *definitions/examples* item showed some evidence of crossloading onto the *Language Complexity* factor (retell M.I. = 28.97; generation M.I. = 51.88). Conceptually, this item should not reflect word- or sentence-level language features. Hence, it is unclear at this time why the item grouped with the *Language Complexity* items for this dataset.

Researchers compared fit indices to evaluate whether model fit improved or got worse as these changes were made. It was discovered that Model 5 displayed the strongest evidence of fit with the data (generation samples $\chi^2(26) = 175.73, p < .001$, CFI = .91, TLI = .87, RMSEA = .08, and SRMR = .05; retell samples $\chi^2(26) = 159.53, p < .001$, CFI = .79, TLI = .71, RMSEA = .07, and SRMR = .05). With the *transition* item removed, the remaining factor pattern loadings for Model 5 model ranged from .27 to .79, with significant values for each item. One indicator, *concluding statement*, loaded onto its respective factor (*Passage Structure*) with values less than

Table 5

Fits of Models That Test Different Conceptualizations of Narrative and Expository Academic Language

Instrument	#	Model Tested	Task Type	Free parameters	Chi-Square		SRMR	CFI	TLI	RMSEA			
					χ^2	df				p -value	RMS EA	90% CI	p $\leq .05$
ELM Flowchart	1	Original model	Generation	31	169.51	34	<.001	0.05	0.91	0.88	0.06	0.053-0.071	0.02
			Retell	31	191.27	34	<.001	0.05	0.79	0.72	0.07	0.058-0.076	0.00
	3	Correlate <i>main idea</i> with <i>passage cohesion</i>	Generation	32	139.35	33	<.001	0.04	0.93	0.90	0.06	0.046-0.065	0.15
			Retell	32	192.32	33	<.001	0.05	0.79	0.71	0.07	0.059-0.078	0.00
	2	Delete main idea	Generation	28	91.67	26	<.001	0.04	0.94	0.92	0.05	0.039-0.061	0.52
			Retell	28	127.11	26	<.001	0.04	0.82	0.75	0.06	0.051-0.072	0.04
	4	Load <i>vocabulary</i> onto <i>Passage Structure</i>	Generation	31	138.42	34	<.001	0.05	0.93	0.91	0.05	0.045-0.064	0.21
			Retell	31	146.91	34	<.001	0.04	0.85	0.80	0.06	0.047-0.066	0.12
	5	Delete <i>transitions</i>	Generation	31	175.73	26	<.001	0.05	0.91	0.87	0.08	0.064-0.085	<.001
			Retell	31	159.53	26	<.001	0.05	0.79	0.71	0.07	0.060-0.081	0.00
NLM Flowchart	1	Original model	Generation	46	743.85	89	<.001	0.06	0.88	0.86	0.08	0.079-0.090	<.001
			Retell	46	784.80	89	<.001	0.05	0.91	0.90	0.09	0.081-0.092	<.001
	2	Correlate <i>problem</i> with <i>plan/attempt</i>	Generation	47	739.93	88	<.001	0.06	0.88	0.86	0.09	0.079-0.090	<.001
			Retell	47	710.23	88	<.001	0.05	0.92	0.91	0.08	0.077-0.088	<.001
	3	Correlate <i>character</i> with <i>setting</i>	Generation	47	727.34	88	<.001	0.05	0.92	0.90	0.08	.078-.089	<.001
			Retell	47	634.52	88	<.001	0.05	0.90	0.88	0.08	.072-.083	<.001

SRMR = standard root-mean-square residual; CFI = comparative fit index; TLI = Tucker Lewis Index; RMSEA = root-mean-square error of approximation

.30 (generation samples 0.27; retell samples 0.28). Factor pattern loadings for the best fitting models of narrative and expository academic language are contained in Table 6.

NLM Flowchart Model Specifications

Modification indices for Model 1 of narrative academic language indicated noteworthy correlations between many pairs of indicators. Correlations between nearly all of the story grammar elements showed up as significant in the modification indices. These correlations were expected, since prior research has described how the three primary story grammar elements (i.e., problem, attempt to solve the problem, and resolution) are causally related (Stein & Glenn, 1979). For the current study, researchers chose to create modified models to control for two indicated correlations: *problem* with *plan/attempt* (Model 2; retell M.I. = 85.41; generational M.I. = non-significant) and *character* with *setting* (Model 3; retell M.I. = 55.65; generation M.I. = 99.54). Changes in fit index values were observed for models reflecting these particular correlations because (a) unusually high M.I. values were observed, and (b) conceptually these concepts are interrelated. Specifically, for Model 2, we reasoned that a *plan/attempt* to solve a problem cannot occur without a *problem* occurring. For Model 3, *character* and *setting* represent background details that are typically the first bits of information presented in a story. Models reflecting other correlations between story grammar elements could be investigated, but for this study we chose to explore these correlations only.

Additionally, there were three items that slightly crossloaded onto contra-indicated factors: *setting* (retell M.I. = 59.58; generation M.I. = 81.08); *character* (retell M.I. = 74.72; generation M.I. = 37.31); and *emotion* (retell M.I. = 80.84; generation M.I. = 35.29).

Conceptually, these items do not fit with the other indicators on the scales they grouped with. It is unclear at this time why these patterns were evident in the data.

Fit indices were evaluated to assess the extent to which model fit improved or got worse as changes were made to the model. Results show the original *NLM Flowchart* model of narrative academic language had the best fit to the data (generation $\chi^2(46) = 743.85, p < .001$, SRMR = .06, RMSEA = .08, CFI = .88, and TLI = .86; retell $\chi^2(46) = 784.80, p < .001$, SRMR = .05, RMSEA = .09, CFI = .91, and TLI = .90). Standardized factor loadings ranged from 0.24 to 0.96, with significant p-values for each item. Two *Narrative Structure* items loaded onto the identified factor with values less than .30: *setting* (generation = 0.29; retell = 0.54) and *emotion* (generation samples .24; retell samples .38). It is important to note that poor loadings factor loadings for these items were below .30 for generation language samples only. Loadings exceeded the cutoff criteria in retell samples.

Intercorrelations between factor structures for the *NLM* and *ELM Flowcharts* are reported in Table 7. Research suggests that intercorrelation values of .80 and below provide sufficient evidence that factors have separate structures and are likely not unidimensional (Brown, 2015). All intercorrelation values between identified factors fell below this threshold.

Reliability of the NLM and ELM Flowcharts

Internal Consistency

Internal consistency reliability of the factors was assessed using Cronbach's alpha. Results are presented in Table 8. Alphas ranged from .40 to .85. Current standards in the research literature identify values greater than or equal to 0.80 as adequate in terms of internal consistency (Nunnally, 1978). The *Narrative Structure* factor of the *NLM Flowchart* (generation = 0.79; retell = 0.85) met this criterion for retell samples only. *Language Complexity* factors for both the *NLM Flowchart* (generation = 0.40; retell = 0.54) and the *ELM Flowchart* (generation = 0.58;

Table 6*Factor Loading Analysis*

Instrument	Factor	Indicator	Generation				Retell			
			Factor Loading	Standard Error	Residual Variance	<i>p</i> -value	Factor Loading	Standard Error	Residual Variance	<i>p</i> -value
NLM Flowchart	Language Complexity	Relative Pronouns	0.375	0.039	0.859	<.001	0.317	0.033	0.900	<.001
		Verb/ Noun Modifiers	0.575	0.030	0.669	<.001	0.742	0.016	0.449	<.001
		Vocabulary/ Rhetoric	0.527	0.038	0.722	<.001	0.721	0.015	0.480	<.001
		Temporal Ties	0.375	0.038	0.860	<.001	0.413	0.030	0.829	<.001
		Causal Ties	0.409	0.035	0.833	<.001	0.416	0.027	0.827	<.001
		Dialogue	0.355	0.035	0.874	<.001	0.488	0.023	0.761	<.001
	<i>Range of Loadings:</i>		<i>.36 - .58</i>				<i>.32 - .74</i>			
	Narrative Structure	Episode Complexity	0.957	0.006	0.085	<.001	0.967	0.004	0.065	<.001
		Character	0.538	0.023	0.710	<.001	0.607	0.024	0.632	<.001
		Setting	0.286	0.038	0.918	<.001	0.535	0.023	0.714	<.001
		Problem	0.774	0.018	0.400	<.001	0.804	0.013	0.353	<.001
		Sequence	0.846	0.014	0.285	<.001	0.873	0.010	0.238	<.001
		Plan/ Attempt	0.848	0.013	0.280	<.001	0.907	0.007	0.177	<.001
		Consequence	0.863	0.012	0.255	<.001	0.909	0.009	0.174	<.001
		Ending	0.528	0.022	0.721	<.001	0.626	0.022	0.609	<.001
		Emotion	0.241	0.031	0.942	<.001	0.377	0.023	0.858	<.001
	<i>Range of Loadings:</i>		<i>.24 - .96</i>				<i>.38 - .97</i>			
ELM Flowchart	Language Complexity	Relative Pronouns	0.533	0.029	0.716	<.001	0.420	0.032	0.823	<.001
		Verb/ Noun Modifiers	0.629	0.026	0.604	<.001	0.628	0.026	0.605	<.001
		Vocabulary	0.646	0.030	0.583	<.001	0.634	0.031	0.598	<.001
		Temporal Ties	0.384	0.039	0.852	<.001	0.362	0.037	0.869	<.001
		Causal Ties	0.608	0.030	0.631	<.001	0.498	0.034	0.752	<.001
	<i>Range of Loadings:</i>		<i>.38 - .65</i>				<i>.36 - .63</i>			
	Passage Structure	Main Idea	0.656	0.022	0.569	<.001	0.549	0.043	0.699	<.001
		Definitions & Examples	0.508	0.792	0.741	<.001	0.416	0.090	0.827	<.001
		Passage Cohesion	0.792	0.031	0.373	<.001	0.780	0.065	0.392	<.001
		Concluding Statement	0.274	0.022	0.925	<.001	0.275	0.044	0.924	<.001
	<i>Range of Loadings:</i>		<i>.27 - .79</i>				<i>.28 - .78</i>			

Table 7*Intercorrelation Estimates*

Instrument	Factors	Task	Intercorrelation	Standard Error	<i>p</i> -Value
ELM Flowchart	Language Complexity x	Generation	0.743	0.038	<.001
	Expository Structure	Retell	0.794	0.054	<.001
NLM Flowchart	Language Complexity x	Generation	0.657	0.032	<.001
	Narrative Structure	Retell	0.80	0.018	<.001

*Standardized factor loadings (STDYX)

Table 8*Cronbach's Alpha Coefficients*

Instrument	Task Type	Cronbach's Alpha α		
		Language Complexity	Narrative Structure	Passage Structure
NLM Flowchart	Generation	0.40	0.79	-
	Retell	0.54	0.85	-
ELM Flowchart	Generation	0.58	-	0.51
	Retell	0.49	-	0.49

retell = 0.49) fell below this threshold. Internal consistency of the *Passage Structure* factor also fell below standards for internal consistency (generation samples 0.51; retell samples 0.49).

Agreement between Raters

Interrater agreement was calculated for each item by dividing the smaller number by the larger number, then multiplying the result by 100%. This value provides an appropriate index of reliability for rate-based measures (Cooper et al., 1987). Cohen's kappa (Cohen, 1960) was calculated to account for the possibility of chance agreement between raters. Reliability data for the *NLM Flowchart* and *ELM Flowchart* are shown in Table 9. Interrater agreement for individual *NLM Flowchart* items ranged between 51% and 96% (mean = 85%). Mean agreement with regard to the *NLM Flowchart* factors was 88% (generation) and 91% (retell) for *Language Complexity*, and 77% (generation) and 78% (retell) for *Narrative Structure*. Interrater agreement

on individual *ELM Flowchart* items ranged between 40% and 99% (mean = 84%). Notably, the *passage cohesion* item was an outlier at 40% (generation) and 42% (retell) agreement; all other items scored 74% and above. *Passage Structure* demonstrated lower levels of agreement (generation mean = 76%; retell mean = 77%) than *Language Complexity* (generation mean = 88%; retell mean = 86%).

Kappa coefficients for two independent raters of items from the *NLM Flowchart* ranged between .39 and .92. The majority of coefficient values suggest substantial agreement between raters, with the exception of four items which displayed weak agreement – *episode complexity* (generation = 0.39; retell = 0.36), *consequence* (generation samples .44; retell samples .41), *problem* (generation = 0.53; retell = 0.51), and *ending* (generation = 0.42; retell = 0.49). Overall, the *NLM Flowchart* demonstrates moderate agreement (generation = 0.66; retell = 0.68). 40% and 99% (mean = 84%). Notably, the *passage cohesion* item was an outlier at 40% (generation) and 42% (retell) agreement; all other items scored 74% and above. *Passage Structure* demonstrated lower levels of agreement (generation mean = 76%; retell mean = 77%) than *Language Complexity* (generation mean = 88%; retell mean = 86%).

Cohen's kappa coefficients for two independent raters of items from the *NLM Flowchart* ranged between .39 and .92. The majority of coefficient values suggest substantial agreement between raters, with the exception of four items which displayed weak agreement – *episode complexity* (generation = 0.39; retell = 0.36), *consequence* (generation samples .44; retell samples .41), *problem* (generation = 0.53; retell = 0.51), and *ending* (generation = 0.42; retell = 0.49). Overall, the *NLM Flowchart* demonstrates moderate agreement (generation = 0.66; retell = 0.68).

Table 9

Interrater Agreement and Coehn's Kappa Values

Instrument	Scale	Item	Retell			Generation		
			% Agreement	Kappa	p-value	% Agreement	Kappa	p-value
<i>NLM Flowchart</i>	Language Complexity	Relative Pronouns	0.89	0.67	<.001	0.93	0.52	<.001
		Verb/Noun Modifiers	0.86	0.74	<.001	0.86	0.80	<.001
		Vocabulary	0.87	0.61	<.001	0.88	0.78	<.001
		Temporal Ties	0.90	0.87	<.001	0.93	0.82	<.001
		Causal Ties	0.86	0.74	<.001	0.89	0.75	<.001
		Dialogue	0.90	0.76	<.001	0.96	0.80	<.001
		<i>Language Complexity</i>	0.88	0.73		0.91	0.75	
	Narrative Structure	Episode Complexity	0.46	0.39	<.001	0.51	0.36	<.001
		Character	0.95	0.88	<.001	0.94	0.91	<.001
		Setting	0.84	0.74	<.001	0.82	0.77	<.001
		Problem	0.80	0.53	<.001	0.85	0.49	<.001
		Sequence	0.89	0.63	<.001	0.85	0.69	<.001
		Plan/Attempt	0.79	0.54	<.001	0.76	0.63	<.001
		Consequence	0.64	0.44	<.001	0.72	0.38	<.001
		Ending	0.75	0.42	<.001	0.68	0.56	<.001
		Emotion	0.85	0.75	<.001	0.86	0.70	<.001
		<i>Narrative Structure</i>	0.77	0.59		0.78	0.61	
		Total	0.83	0.66		0.78	0.68	
	Language Complexity	Relative Pronouns	0.92	0.68	<.001	0.87	0.73	<.001
		Verb/ Noun Modifiers	0.77	0.57	<.001	0.74	0.59	<.001
		Vocabulary	0.87	0.84	<.001	0.90	0.80	<.001
		Temporal Ties	0.97	0.86	<.001	0.93	0.90	<.001
		Casual Ties	0.88	0.76	<.001	0.86	0.77	<.001
		<i>Language Complexity</i>	0.88	0.74		0.86	0.76	
	Passage Structure	Main Idea	0.88	0.33	<.001	0.83	0.46	<.001
		Definitions & Examples	0.79	0.51	<.001	0.87	0.30	<.001
		Passage Cohesion	0.40	0.08	<.001	0.42	0.11	<.001
		Concluding Statement	0.96	0.14	<.001	0.96	0.00	<.001
		<i>Passage Structure</i>	0.76	0.27		0.77	0.22	
		Total	0.82	0.50		0.82	0.49	

Kappa coefficients for the *ELM Flowchart* ranged between .00 and .90. All items from the *Language Complexity* factor displayed moderate to strong coefficient values. In contrast, kappas for every item contained in the *Passage Structure* factor were at or below .40, with the exception of *definition and examples* (generation = 0.51; retell = 0.30). These kappa coefficients suggest that, after controlling for chance agreement, the *Passage Structure* factor displays low to very low reliability.

Correlations with Other Measures

A bivariate correlation analysis was conducted in SPSS to determine correlations between scores from the *WJ-IV TOL* and the *NLM* and *ELM Flowcharts*, respectively. A small, positive correlation was observed between *WJ-TOL CALP* scores and the *NLM* (generation = 0.22; retell = 0.29) and *ELM Flowcharts* (generation = 0.26; retell = 0.25). Correlation values and 95% confidence intervals are reported in Table 10.

Table 10

Correlations With WJ-TOL CALP Scores

	Factor	Generation		Retell	
		<i>r</i>	95% CI	<i>r</i>	95% CI
NLM Flowchart	Language Complexity	0.144	.100-.187	0.238	.195-.279
	Narrative Structure	0.21	.167-.251	0.283	.242-.323
	Total	0.221*	.178-.262	0.294**	.253-.334
ELM Flowchart	Language Complexity	0.219	.177-.260	0.205	.162-.247
	Passage Structure	0.247	.205-.288	0.215	.172-.257
	Total	0.264***	.223-.304	0.249****	.207-.291

* $n = 1,966$; ** $n = 1,966$; *** $n = 2,008$; **** $n = 1,947$

CHAPTER 5: DISCUSSION

The purpose of this study was to examine the psychometric properties of two novel discourse analysis tools designed to measure the spoken academic language of children in kindergarten through third grade. Unlike current methods of direct academic language measurement which focus on elements of language in isolation (e.g., microstructure and macrostructure), the *NLM* and *ELM Flowcharts* enable a direct assessment of spoken language that is both comprehensive and discourse-specific. The present study aimed to determine the extent to which these instruments can generate dependable, accurate information about spoken academic language. 7,887 language samples derived from a previous cohort-design study of K-3rd grade students ($n = 1,040$) from different racial/ethnic, SES, and family language backgrounds were scored.

Overall, the distribution of item-level data generated by the *Flowcharts* did not follow a strictly normal trend. It was expected that ceiling and floor effects would impact item distributions because of the narrow range of possible values for *Flowchart* items. We used an estimator that is robust to non-normal multivariate distributions (WLSMV) for the CFA to accommodate for this trend in the data.

Data distributions varied by instrument. *Language Complexity* mean scores were higher for expository samples than for narrative samples. This finding is not surprising, given that there are well-documented differences between genres favoring expository structures in terms of increased linguistic complexity (Schleppegrell, 2001). *Narrative Structure* score means were

higher than *Passage Structure* means, which was an expected trend. Narrative discourse production and comprehension develop well before equivalent skills develop in expository discourses (Lundine et al., 2018). Hence, K-3rd grade students should not be expected to perform equitably in both discourse genres.

For both the *NLM* and the *ELM Flowchart*, individual item mean scores were inconsistent across task types. Mean scores for *NLM Flowchart* generation tasks were, on average, lower than mean scores for *NLM Flowchart* retell tasks. In other words, students displayed more academic features in their language when retelling a story than when asked to generate a story of their own making. This finding is in line with prior research. For both language-impaired and typically developing children, narrative retells tend to be longer than narrative generations, include more story grammar components, and more complete episodic structures (Merritt & Liles, 1989). In contrast, mean scores for *ELM Flowchart* generation tasks were higher than for retell tasks. This finding conflicts somewhat with a previous report (Westerveld & Moran, 2013) in which expository language samples were elicited from primary, middle, and high school students in both retell and generation contexts. In the study, language samples from the retell condition were significantly longer and more complex (e.g., higher clausal density) than those elicited in generation conditions. However it should be noted that only microstructural elements were assessed. Moreover, a single stimulus was administered for the retell task (i.e., a retell passage about the game of curling) and for the generation task (i.e., the verbal question, “what is your favorite game or sport, and why?). Hence, findings from this study may be less reflective of the academic language construct and more reflective of these methodological limitations. The *ELM Flowchart*, which attempts to directly measure discourse features, may provide a more comprehensive estimate of students’ overall expository academic language ability than proxy

estimates derived from microstructural analyses. Furthermore language scores generated through the *ELM Flowchart* may be more generalizable, since a wider selection of stimuli were used to elicit both generation and retell language samples.

One explanation for the differential performance across genres is that students may integrate and reproduce complex language features better when they are presented in narrative form. There are several differentiating characteristics between genres that may make complex language features more or less difficult to understand (Schleppegrell, 2001; Lundine & McCauley, 2016; Snow et al., 2010). First, expository language tends to vary widely by form and function, whereas the narrative form is typically predictable and uniform (Duke et al., 2011). Second, narrative elements connect with each other through a series of largely predictable causal and temporal relations (Turner, 1996). In contrast, expository elements (i.e., information units) do not connect in highly predictable ways, making it more challenging for readers to create meaning from the text (Hill et al., 2021). Furthermore, expository texts are more likely to contain domain-specific, specialized vocabulary. Young students may not yet have the content knowledge required to interpret novel terminology and integrate it into their retell responses (Schleppegrell, 2001). For these reasons, complex language features may be less accessible to young students when presented in expository form. If this is the case, then students would be expected to struggle more with reproducing the complex language heard in expository retell tasks than they would with forming a self-generated informational text.

Measurements of Reliability

The *NLM* and *ELM Flowcharts* showed mixed evidence of interrater reliability across subscales. Findings will be discussed separately for each subscale, along with interpretations

based on theory and prior research. With the exception of the *NLM Flowchart Narrative Structure* subscale, internal consistency reliability did not meet benchmarks of acceptability. A discussion of how to interpret these findings through a formative conceptual model follows.

Interrater Agreement

Language Complexity Subscale. Scorers achieved moderate levels of agreement on *Language Complexity* composites across genre and task type. Individual items from the *Language Complexity* subscale demonstrated interrater agreement scores ranging from weak (i.e., *relative pronouns* and *verb/noun modifiers*) to moderate. These findings were consistent even after controlling for chance agreement via Cohen's kappa.

Narrative Structure Subscale. With the exception of *episode complexity*, *consequence*, and *ending*, adequate rates of agreement between scorers were found for each *Narrative Structure* indicator. Low agreement on the *consequence* and *ending* items is likely due to these constructs being conceptually related and difficult to distinguish in actual student language (Peterson, 1990). The *episode complexity* item represents a summative score that is dependent upon the presence of other items. For example, a language sample with a score of 3 for *problem* and a score of 3 for *plan/attempt* would receive an *episode complexity* score of 2. If scorers disagreed by one point on the *problem* item, they would also disagree on *episode complexity*; hence, the odds of disagreement between raters are much higher for this item than for other items in the *Flowchart*.

When averaged together, agreement values for all of the original *Narrative Structure* items evidenced acceptable rates of agreement across both generation and retell samples. This finding held constant even after controlling for chance agreement via Cohen's kappa.

Passage Structure Subscale. Point-by-point percent agreement scores for all *Passage Structure* items were within the acceptable range (75% or above) across task types, with the exception of the *passage cohesion* item. However, after accounting for chance agreement via Cohen's kappa, *none* of the items were found to have been rated consistently across task types. These findings suggest that the *Passage Structure* items require further refinement in order to increase the level of agreement between raters.

Internal Consistency

Alphas for the retell samples of the *NLM Flowchart Narrative Structure* subscale exceeded standards of acceptability (0.85). The alpha coefficient for generation samples was only .01 beneath the cut-off (0.79). Hence, it can be said that the internal structure of the *NLM Flowchart Narrative Structure* subscale is somewhat consistent. In contrast, low Cronbach's alpha coefficients were found across task types for the *Language Complexity* and *Passage Structure* subscales, suggesting that the items they contain are not closely related as a group. Theoretical models impact how data should be analyzed and interpreted (Bollen, 2011). Therefore, the precise nature of the conceptual model guiding our interpretation of academic language requires further consideration.

Conceptually, there are many aspects of the academic language construct that we perceive as being formative in nature. According to Bollen (2011), there are important distinctions between formative and reflective measurement models which influence how they operate. Conceptually, the latent variable(s) of reflective models can be said to exert some kind of influence or effect on certain indicators. Hence, these are typically referred to as "effect indicators". Since effect indicators are all directly reflective of the latent variable, any indicator

can be selected, substituted or deleted and the construct will still be left intact. In contrast, formative model indicators are referred to as “causal indicators” because *together* they form (i.e., cause) the latent variable. The latent variable of a formative measure can be thought of as a “useful summary device for the effect of several variables on other variables” (Bollen, 2011, p. 360). Each variable (i.e., indicator) contributes a slightly *different* aspect to the overall construct. Hence, it is not reasonable to expect causal indicators to group perfectly together.

Of course, the internal consistency of any measurement model can theoretically be improved by deleting divergent indicators, and this is often the standard approach to improving reliability values for reflective measures (Tavakol & Dennick, 2011). However, for the reasons cited above, the deletion approach is not always recommended for formative measurement models (Bollen, 2011). In fact, measurement experts have cautioned against using effect indicator selection tools (e.g., Cronbach’s alpha, item-total correlations, etc.) to make decisions about causal indicators (Bollen & Lennox, 1991). We theorize that the sub-standard internal consistency values identified for the *Language Complexity* and *Passage Structure* subscales should be interpreted through this lens. Further discussion of this concept will be explored further on in this report.

Measurements of Validity

The *NLM* and *ELM Flowcharts* showed mixed evidence of validity. Findings will be discussed separately for each instrument, with interpretations based on theory and prior research.

NLM Flowchart

The *NLM Flowchart* in its original form showed evidence satisfying multiple conditions to support the instrument’s validity, with regard to its intended purpose of generating accurate,

useful data to better inform the development of empirically-based, narrative academic language interventions. This study revealed that on average, students' *NLM Flowchart* scores increase as they progress through primary school, indicating that the instrument may be sensitive to changes associated with progressive language development. Using a single-level confirmatory factorial design, this study also provided evidence that at least two latent abilities are responsible for children's performance on the *NLM Flowchart*. In accordance with the original model, narrative academic language can be approximated by the items contained within the *Language Complexity* and *Narrative Structure* subscales. Admittedly, Model 1's fit index values were not very robust; however, they exceeded the values of four alternative models, and passed the test of fit originally defined by this research team.

Poor factor loadings were noted for *setting* and *emotion* with regard to generation samples only; these items loaded moderately for retell samples. It is not entirely clear why *setting* and *emotion* were not significant contributors to the *Narrative Structure* factor in the generation language samples. Prior research indicates that children as young as 3-years-old are able to make inferences about the internal states of story characters (Deconti and Dickerson, 1994) and that children age four through six generally include setting in their retelling of stories (Stein & Glenn, 1979). Some degree of prompting or scaffolding may be needed in order for children to produce these features without the aid of a retell model.

Model 1, which contained both *setting* and *emotion* indicators, was found to fit best with the data for both generation and retell samples. Modification indices did not suggest that removing either of these items would significantly improve model fit; however, the effects of deleting these indicators from the instrument were not evaluated in the course of this study. Future researchers may wish to evaluate how the overall *NLM Flowchart* factor structure changes when

these items are deleted from the list of factor indicators. However, there are significant limitations associated with the deletion approach to assessment development, especially for formative measurement models. A discussion of these limitations will be addressed near the end of this chapter.

Finally, this study found a weak positive correlation between student scores from the *NLM Flowchart* and scores from the *Woodcock-Johnson IV Test of Oral Language*. Student performance on standardized language tests such as the *WJ-IV TOL* are influenced by a host of factors that are related to, but not directly reflective of their ability to produce and comprehend language (McNamara, 2001). Standardized tests require skills in inferencing, short term memory processes, and sustained attention, to name just a few. This makes sense, given that standardized test batteries such as the *WJ-IV* are designed to be used in tandem to measure broad-ranging cognitive or academic attributes (Dombrowski et al., 2019). In contrast, the specificity of the *NLM Flowchart* more directly measures the skills uniquely associated with speaking and understanding academic language. Furthermore, the *NLM Flowchart* focuses specifically on *narrative* academic language, which has been shown to be unique in form and function (Stein & Glenn, 1979). A weak correlation between scores from the *NLM Flowchart* and the *WJ-IV TOL* may be reflective of these important differences between the two instruments.

ELM Flowchart

Concerning validity of the *ELM Flowchart*, this study found that two of three conditions for support were satisfied. First, on average students' *ELM Flowchart* scores increased across grade levels, indicating that the instrument may reflect developmental trends in language development. Second, a single-level CFA indicated that the original two-factor model failed to

adequately describe children's performance on the *ELM Flowchart*. The *transitions* and *concluding statements* items demonstrated poor factor loading onto the *Language Complexity* and *Passage Structure* factors, respectively. These two items explained less than 10% of the overall variance for their respective factors.

Readers may conclude that these two items do not contribute to the latent variables in a meaningful way for this population, and should therefore be removed from the instrument. Indeed, this study tried deleting the *transitions* item and found that with the specified model, overall *ELM Flowchart* model fit reached an acceptable range. Nevertheless, we caution against the deletion approach and instead suggest that future attention to the *ELM Flowchart* should be given to refining the current items, and selecting additional indicators to enhance the model. Practical and conceptual reasons for this alternative approach are outlined more fully in the following section of this report.

There was some evidence that the *vocabulary* item may cross-load onto both factors. Language features such as vocabulary tend to overlap, and are difficult to distinguish in spoken language (Petersen, 2011). In Model 4, we grouped *vocabulary* under the *Passage Structure* factor and found that model fit was not significantly enhanced. Hence it is our conclusion that *vocabulary* fits sufficiently well with the other *Language Complexity* items and should remain within that factor.

Finally, this study found a weak positive correlation between student scores from the *ELM Flowchart* and a standardized, norm-referenced assessment of academic language. Like the *NLM Flowchart*, the *ELM Flowchart* was designed to be specifically informative about the unique skills associated with speaking and understanding expository academic language. Hence, a weak correlation between scores from the *ELM Flowchart* and the *WJ-IV TOL* may be

reflective of important differences in terms of assessment type (direct vs. indirect) and/or the variable of interest (academic language vs. expository academic language).

Conclusions, Implications, and Future Directions

Academic language, or the “language of schooling”, is a constellation of distinct word, sentence, and discourse level patterns (Schleppegrell, 2001). Academic language skills are thought to be vital to students’ academic development in all subject areas. Moreover, research has demonstrated that academic language proficiency in early childhood is strongly predictive of reading comprehension in later childhood (Uccelli et al., 2015). High-quality measures of academic language are necessary to better understand how the construct functions and how it develops over time. Extant methods of assessing academic language (e.g., norm-referenced tests; structural assessments) are insufficient for the task of understanding its varied features. Due to the importance of academic language for literacy and learning, better tools for measuring and analyzing academic language are needed.

The current research study presented an in-depth psychometric analysis of two instruments designed to assess the spoken academic language of primary-grade students. Genre-specific academic language was scored along two dimensions: *Language Complexity* (i.e., lexical/grammatical microstructure) and *Narrative/Passage Structure* (i.e., discourse-specific macrostructure).

NLM Flowchart

This study provides further evidence that the *NLM Flowchart* reliably and accurately measures spoken, narrative academic language across two dimensions – *Language Complexity*

(i.e., lexical/grammatical microstructure) and *Narrative Structure* (i.e., discourse-specific macrostructure). Our findings are consistent with the results of prior studies which document the *NLM Flowchart*'s ability to reliably track progressive changes in young students' oral language (Spencer et al., 2013; Petersen & Spencer, 2019). Our results support the claim that the *NLM Flowchart* subscales can be used independently or in tandem to assess the productive spoken language of students in kindergarten through 3rd grade along these dimensions. The instrument successfully produced reliable, useful data about language samples elicited through story retell or story generation tasks. These findings support use of the *NLM Flowchart* for research applications to better understand, for example, how academic language varies with respect to student characteristics, language elicitation contexts, etc.

Future research on the *NLM Flowchart* should focus on readying the instrument for use in applied settings. For example, the *NLM Flowchart* would be beneficial in school settings where a response-to-intervention (RtI) framework is employed to distinguish children with language disorders from typically-developing peers, and/or to inform decisions about language interventions. Next steps should include (1) establishing sensitivity of the *NLM Flowchart* to intervention effects; (2) conceptualizing a method for measuring students' responsiveness to instruction using the *Flowchart*, and (3) establishing a criterion for defining non-responsiveness (Fuchs & Fuchs, 2006).

Future research may also investigate more deeply the internal structure of the *NLM Flowchart* to explore how different iterations of the model might change fit index values. In this study, we controlled for only two correlations between *Narrative Structure* items, even though there were many more correlations noted in the modification indices. Running comparative fit

indices (e.g., AIC, BIC) would be a valuable approach to providing more direct comparisons between these models.

ELM Flowchart

In contrast to the *NLM Flowchart*, this study presented only preliminary findings regarding the psychometric properties of the *ELM Flowchart*, a completely novel instrument. We found that raters were not consistent in their scoring of the four selected items from the *ELM Flowchart Passage Structure* subscale (i.e., *main idea*, *passage cohesion*, *definitions/examples*, and *concluding statement*). It may be the case that scoring ambiguities resulted from the instrument being tested on young children in whom these language structures are not yet developed. Future research should definitely explore applications of the *ELM Flowchart* with academic language sampled from students in the upper primary (third through fifth) and middle (sixth through eighth) grades.

Nonetheless, until interrater agreement values for these items are improved through, for example, refinement of the operational definitions and training procedures, we cannot suggest using the *Passage Structure* subscale to measure expository discourse features in the productive language of young children. We expect that any modifications resulting from such efforts would significantly impact variable distributions and relationships. Hence, factor analyses would need to be repeated to determine the structure of the resultant data.

Our findings suggest that the *Language Complexity* subscale of the *ELM Flowchart* in its current form can be reliably administered by multiple raters. When the *transitions* item is removed from the assessment, this subscale can be said to generate an accurate representation of the lexical/grammatical features of young children's spoken expositions, whether elicited

through passage retell or passage generation tasks. However, caution should be exhibited with regard to removing the *transition* item from the *Flowchart*. There are several limitations to the deletion approach that will be discussed in the section that follows. A better option would be to identify additional indicators that might contribute to the structure of the *Language Complexity* factor. As per our previous comments, this can be said for the *Passage Cohesion* factor, as well.

Some researchers may be interested in using the *ELM Flowchart* in its current form, despite the significant limitations described here. We wish to emphasize that at this point, the instrument is an inconsistent “use at your own risk” tool that may or may not generate useful information about expository academic language in general, and expository discourse features in particular. It is worth noting that the *ELM Flowchart* in its current form includes a summative item measuring information units. Information units is a language variable that features prominently in research on expository academic language (Black, 2017). For this study, it was decided that the information units indicator should be excluded from the factor analysis because it was anticipated to contribute too much variance to the identified factor, thereby obscuring the variance attributable to other variables. Future studies should investigate how the internal structure of the *Passage Structure* subscale changes when this item is added to the battery of refined indicators. For a more immediate workaround, one might consider replacing the four *Passage Structure* items with *information units*, and use that item in isolation as a rough estimate of expository macrostructure.

Support for Finding Alternatives to Item Deletion

In line with current CFA item selection practices, future researchers may be tempted to delete items with poor factor loadings from the *Flowcharts* in an effort to increase internal consistency

reliability and other psychometric indices (Mueller & Hancock, 2001). There are noted limitations to the deletion approach that warrant explicit discussion, several of which apply specifically to formative measurement models. The acceptability of making post hoc changes (e.g., item deletions) to models hypothesized and tested through a CFA is highly debated. As Bandalos and Finney (2010) note, “Researchers must keep in mind that the purpose of conducting a CFA study is to gain a better understanding of the underlying structure of the variables, not to force models to fit” (p. 112). Hence, future studies would need to test any modified *Flowchart* models with new datasets, to ensure that the model is not being forced to fit with the data it is being tested on.

Secondly, removing items that do not fit well with other items places limitations on an instrument’s ability to monitor progress over time. In the case of the *Flowcharts*, these items may be representative of language features that young children have not yet learned to use, but are nonetheless important contributors to the developing academic language construct. These items were selected based on their alignment with academic standards, as well as their well-established documentation in the academic language research literature. In the current study, fewer students overall were able to produce the language structures captured by *Flowchart* items with poor factor loadings. Hence, it may be important to retain these items so that the language growth of older (e.g., 5th – 8th grade) students can be determined.

It is the opinion of the researchers that future research on the *ELM Flowchart* should look for alternatives to deleting items. The first potential alternative suggested by our findings would be to simply add more indicators to the respective subscales. A potential outcome of such efforts might be significant increases in internal consistency values. The latent variables (subscales) included in the *NLM* and *ELM Flowcharts* were derived from a rich body of research literature in

the speech/language cannon, which clearly distinguishes between microstructural (lexical/grammatical) and macrostructural (discursive) language components (Petersen, 2011). Drawing from this body of knowledge, indicators were selected for each of the identified variables. There are many additional variables that could be included in the *Flowcharts* which we would expect to contribute to the latent variables in the model. For example, microstructural elements such as MLU, T-UNITS, and clausal density could be added to the *Language Complexity* subscale to achieve this end. Nevertheless, it should be understood that the scope of the academic language construct is extremely broad, multidimensional, and difficult to capture. There is a conflict between instrument specificity, comprehensiveness, and usability that makes accurate, useful language assessments challenging to design and interpret. There is a limit to the number of items that can be included in an instrument for which brevity and usability are among the end goals of development (Lewis et al., 2015).

A second alternative suggested by our findings would be to apply psychometric procedures and inclusion/exclusion criteria better suited for formative measurement models. For example, a collinearity assessment, redundancy analysis, or robustness check may produce more valuable information about the structure and content of the *NLM and ELM Flowcharts* than traditional CFAs (Ghasemy, 2021). Within a formative model, there must be a sufficient “census” of causal indicators to accurately capture the true form of the latent variable (Bollen & Lennox, 1991, p. 307-308). In the case of constructs that are more formative in nature, adding more contributing indicators would be expected to create a more comprehensive formation of the latent variable. Deleting variables, on the other hand, may compromise the integrity of the construct as a whole. Removing any of these four items would likely be detrimental to the overall accuracy of the instruments in their representation of the latent variables.

Limitations of the Current Study

The findings reported in this study are for K-3rd grade students in Florida schools that volunteered to participate. The extent to which findings generalize to other grade levels, schools, or regions is tenuous. The dataset for this study included spoken language samples generated through standardized language elicitation procedures. While our research indicates that the procedures employed are best practice for eliciting high-quality oral language samples, they may also limit generalizability of our findings. Therefore, there is a need for future research to employ the *NLM* and *ELM Flowcharts* with oral and written language samples generated through different elicitation contexts.

Finally, there is a chance that there may be a different latent variable that explains the structures uncovered in this study. For example, items contained in the *NLM* and *ELM Flowcharts* may load onto generalized intelligence (G) factor, or language broadly, rather than discourse-specific academic language constructs specifically. There is such little research on the dimensions of academic language and their measurement that additional latent variables can and should be explored in future research, especially as more academic language research is conducted.

REFERENCES

- Adlof, S. M., & Hogan, T. P. (2019). If we don't look, we won't see: Measuring language development to inform literacy instruction. *Policy Insights from the Behavioral and Brain Sciences*, 6(2), 210-217. <https://doi.org/10.1177/2372732219839075>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.
- Bandalos, D. L., & Finney, S. J. (2010). Exploratory and confirmatory factor analysis. In D.L Bandalos, S.J. Finney, G.R. Hancock & R.O Mueller (Eds.) *The reviewer's guide to quantitative methods in the social sciences*. New York, Routledge.
- Best, R. M., Floyd, R. G., & Mcnamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology*, 29(2), 137-164.
- Beytollahi, S., Soleymani, Z., & Jalaie, S. (2020). The development of a new test for consecutive assessment of narrative skills in Iranian school-age children. *Iranian Journal of Medical Sciences*, 45(6), 425.
- Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. *Language and Computers*, 26, 181-190.
- Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?. *TESOL Quarterly*, 45(1), 5-35.
- Black, J. B. (2017). An exposition on understanding expository text. In B.K. Britton & J.B. Black (Eds.) *Understanding expository text* (pp. 249-267). Routledge.
- Bollen, K.A. (2011). Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly*, 35(2), 359-372.
- Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, 110 (2), 305.
- Bowles, R. P., Justice, L. M., Khan, K. S., Piasta, S. B., Skibbe, L. E., & Foster, T. D. (2020). Development of the Narrative Assessment Protocol-2: A tool for examining young children's narrative skill. *Language, Speech, and Hearing Services in Schools*, 51(2), 390-404.

- Brown, T.A. (2015). *Confirmatory factor analysis for applied research* (2nd Edition). The Guilford Press.
- Cahill, P., Cleave, P., Asp, E., Squires, B., & Kay-Raining Bird, E. (2020). Measuring the complex syntax of school-aged children in language sample analysis: A known-groups validation study. *International Journal of Language & Communication Disorders*, 55(5), 765-776.
- Cain, K., Oakhill, J. V., Barnes, M. A., & Bryant, P. E. (2001). Comprehension skill, inference-making ability, and their relation to knowledge. *Memory & Cognition*, 29(6), 850-859.
- Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, 96(1), 31.
- Cass, C. E. (1999). The Kaufman survey of early academic and language skills (K-SEALS). *Diagnostic*, 24(1-4), 135-144.
- Cervetti, G. N., Pearson, P. D., Palincsar, A. S., Afflerbach, P., Kendeou, P., Biancarosa, G., ... & Berman, A. I. (2020). How the reading for understanding initiative's research complicates the simple view of reading invoked in the science of reading. *Reading Research Quarterly*, 55, S161-S172.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37-46.
- Common Core State Standards Initiative. (2010). Common core state standards for English language arts & literacy in history/social studies science, and technical subjects.
- Common Core State Standards Initiative. (2021, September 4). *Key shifts in English language arts*. Common Core State Standards Initiative. <http://www.corestandards.org/other-resources/key-shifts-in-english-language-arts/>
- Cooper, J. O., Heron, T. E., & Heward, W. L. (1987). *Applied behavior analysis*. Columbus, OH: Merrill.
- Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimal age question, and some other matters. *Working Papers on Bilingualism*, 19, 197-205.
- Curenton, S. M. (2011). Understanding the landscapes of stories: The association between preschoolers' narrative comprehension and production skills and cognitive abilities. *Early Child Development and Care*, 181(6), 791-808.
- Curenton, S.M., & Justice, L.M. (2004). African American and Caucasian preschoolers' use of decontextualized language: Literate language features in oral narratives. *Language, Speech, and Hearing Services in Schools*, 35, 240-253.

- Deconti, K. A. & Dickerson, D. J. (1994). Preschool children's understanding of the situational determinants of others' emotions. *Cognition and Emotion*, 8, 453–472.
- Dickinson, D. K., & Tabors, P. O. (2002). Fostering language and literacy in classrooms and homes. *Young Children*, 57(2), 10-19.
- Dombrowski, S. C., Beaujean, A. A., McGill, R. J., & Benson, N. F. (2019). The Woodcock-Johnson IV tests of achievement provides too many scores for clinical interpretation. *Journal of Psychoeducational Assessment*, 37(7), 819-836.
- Duke, N.K., Pearson, P.D., Strachan, S.L., & Billman, A.K. (2011). Essential elements of fostering and teaching reading comprehension. In S.J. Samuels & A.E. Farstrup (Eds.), *What research has to say about reading instruction* (4th ed., pp. 51– 93). Newark, DE: International Reading Association.
- Farnia, F., & Geva, E. (2013). Growth and predictors of change in English language learners' reading comprehension. *Journal of Research in Reading*, 36(4), 389-421.
- Fey, M. E., Catts, H. W., Proctor-Williams, K., Tomblin, J. B., & Zhang, X. (2004). Oral and written story composition skills of children with language impairment. *Journal of Speech, Language, and Hearing Research*, 47, 1301–1318.
- Finestack, L. H., Payesteh, B., Disher, J. R., & Julien, H. M. (2014). Reporting child language sampling procedures. *Journal of Speech, Language, and Hearing Research*, 57(6), 2274-2279.
- Foorman, B. R., Petscher, Y., Stanley, C., & Truckenmiller, A. (2017). Latent profiles of reading and language and their association with standardized reading outcomes in kindergarten through tenth grade. *Journal of Research on Educational Effectiveness*, 10(3), 619-645.
- Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it?. *Reading Research Quarterly*, 41(1), 93-99.
- Garcia-Bonery, L. (2011). *The relationship between cognitive academic levels of proficiency and response to intervention tier assignment and the implications for special education* (Publication No. 550203). [Doctoral dissertation, Texas Woman's University]. ProQuest Dissertations & Theses Global.
- Geva, E., & Farnia, F. (2012). Developmental changes in the nature of language proficiency and reading fluency paint a more complex view of reading comprehension in ELL and EL1. *Reading and Writing*, 25(8), 1819-1845.
- Gillam, S. L., Fargo, J., Petersen, D. B., & Clark, M. (2012). Assessment of structure dependent narrative features in modeled contexts: African American and European American children. *English Linguistics Research*, 1(1), 1-17.
- Gillam, S. L., Gillam, R. B., Fargo, J. D., Olszewski, A., & Segura, H. (2017). Monitoring indicators of scholarly language: A progress-monitoring instrument for measuring narrative discourse skills. *Communication Disorders Quarterly*, 38(2), 96-106.

- Gillam, R. B., & Pearson, N. (2017). *Test of Narrative Language*. Austin, TX: PRO-ED.
- Goodwin, A. P., & Ahn, S. (2010). A meta-analysis of morphological interventions: Effects on literacy achievement of children with literacy difficulties. *Annals of Dyslexia*, 60(2), 183-208.
- Gottlieb, M., & Ernst-Slavit, G. (2014). *Academic language in diverse classrooms: Definitions and contexts*. Corwin Press.
- Gough, P. B., & Turner, W. E., (1986). Decoding, reading, and reading disability. *Remedial and Special Education (RASE)*, 7(1), 6-10.
- Granados, A., & Lorenzo, F. (2021). English L2 connectives in academic bilingual discourse: a longitudinal computerised analysis of a learner corpus. *Revista Signos, Estudios de Lingüística*, 54(106).
- Gummersall, D. M., & Strong, C. J. (1999). Assessment of complex sentence production in a narrative context. *Language, Speech, and Hearing Services in Schools*, 30(2), 152-164.
- Guo, L. Y., Schneider, P., & Harrison, W. (2021). Clausal density between ages 4 and 9 years for the Edmonton Narrative Norms Instrument: Reference data and psychometric properties. *Language, Speech, and Hearing Services in Schools*, 52(1), 354-368.
- Hadley, P. A. (1998). Language sampling protocols for eliciting text-level discourse. *Language, Speech, and Hearing Services in Schools*, 29(3), 132-147.
- Hakuta, K., Butler, Y.G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* Santa Barbara, CA: University of California Linguistic Minority Research Institute
- Halliday, M. A. (1993). Towards a language-based theory of learning. *Linguistics and Education*, 5(2), 93-116.
- Hayward, D. V., Stewart, G. E., Phillips, L. M., Norris, S. P., & Lovell, M. A. (2008b). At-a-glance test review: Test of narrative language (TNL). In D. Hayward, E. Stewart, L.M. Phillips & S.P Norris (Eds.) *Language, phonological awareness, and reading test directory* (pp. 1-3). Edmonton, AB: Canadian Centre for Research on Literacy.
- Hill, E., Whitworth, A., Boyes, M., Ziegelaar, M., & Claessen, M. (2021). The influence of genre on adolescent discourse skills: Do narratives tell the whole story?. *International Journal of Speech-Language Pathology*, 23(5), 475-485.
- Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, 2(2).
- Hu, L., Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1-55.

- Hutchinson, J. M., Whiteley, H. E., Smith, C. D., & Connors, L. (2003). The developmental progression of comprehension-related skills in children learning EAL. *Journal of Research in Reading*, 26(1), 19-32.
- Johnston, J. R. (2008). Narratives: Twenty-five years later. *Topics in Language Disorders*, 28(2), 93-98.
- Justice, L. M., Bowles, R., Pence, K., & Gosse, C. (2010). A scalable tool for assessing children's language abilities within a narrative context: The NAP (Narrative Assessment Protocol). *Early Childhood Research Quarterly*, 25(2), 218-234.
- Kaufman, A. S. (1993). *Kaufman Survey of Early Academic and Language Skills: K-SEALS*. American Guidance Service.
- Kirby, M. S., Spencer, T. D., & Chen, Y. J. I. (2021). Oral narrative instruction improves kindergarten writing. *Reading & Writing Quarterly*, 1-18.
- Laija-Rodríguez, W., Ochoa, S. H., & Parker, R. (2006). The crosslinguistic role of cognitive academic language proficiency on reading growth in Spanish and English. *Bilingual Research Journal*, 30(1), 87-106.
- LaRusso, M., Kim, H. Y., Selman, R., Uccelli, P., Dawson, T., Jones, S., ... & Snow, C. (2016). Contributions of academic language, perspective taking, and complex reasoning to deep reading comprehension. *Journal of Research on Educational Effectiveness*, 9(2), 201-222.
- Law, K. S., Wong, C. S., & Mobley, W. M. (1998). Toward a taxonomy of multidimensional constructs. *Academy of Management Review*, 23(4), 741-755.
- Lesaux, N. (2006). Building consensus: Future directions for research on English language learners at risk for learning difficulties. *Teachers College Record*, 108(11), 2406-2438.
- Lewis, C. C., Fischer, S., Weiner, B. J., Stanick, C., Kim, M., & Martinez, R. G. (2015). Outcomes for implementation science: An enhanced systematic review of instruments using evidence-based rating criteria. *Implementation Science*, 10(1), 1-17.
- Lundine, J. P., & McCauley, R. J. (2016). A tutorial on expository discourse: Structure, development, and disorders in children and adolescents. *American Journal of Speech-Language Pathology*, 25(3), 306-320.
- Lundine, J. P., Harnish, S. M., McCauley, R. J., Blackett, D. S., Zezinka, A., Chen, W., & Fox, R. A. (2018). Adolescent summaries of narrative and expository discourse: Differences and predictors. *Language, Speech, and Hearing Services in Schools*, 49(3), 551-568.
- Lundine, J. P. (2020). Assessing expository discourse abilities across elementary, middle, and high school. *Topics in Language Disorders*, 40(2), 149-165.
- Ghasemy, M. (2021, February 10). *Formative measurement model evaluation uipdated* [Video]. YouTube. <https://www.youtube.com/watch?v=Vz793CsxJ5A>

- McCabe, A., & Peterson, C. (1984). What makes a good story. *Journal of Psycholinguistic Research*, 13(6), 457-480.
- McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, 18(4), 333-349.
- Merritt, E.E. & Liles, B.Z. (1989). Narrative analysis: Clinical applications of story generation and story retelling. *Journal of Speech and Hearing Disorders*, 54(3), 438-447.
- Meyer, B. J., Young, C. J., & Bartlett, B. J. (2014). *Memory improved: Reading and memory enhancement across the life span through strategic text structures*. Psychology Press.
- Miller, J. F., & Iglesias, A. (2008). *Systematic Analysis of Language Transcripts (SALT), English & Spanish (Version 9)* [Computer software]. Madison, WI: University of Wisconsin-Madison, Waisman Center, Language Analysis Laboratory.
<http://www.languageanalysislab.com/>
- Miller, J. F., Gillon, G., & Westerveld, M. (2010). *Systematic Analysis of Language Transcripts (SALT), New Zealand Version 2010* [computer software]. Madison, WI: SALT Software LLC.
- Mosenthal, P. B. (1985). Defining the expository discourse continuum: Towards a taxonomy of expository text types. *Poetics*, 14(5), 387-414.
- Mueller, R.O. & Hancock, G.R. (2001). Factor analysis and latent structure, confirmatory. In N.J. Smelser & P.B. Baltes (Eds.) *International Encyclopedia of the Social and Behavioral Sciences* (pp. 5239-5244) Oxford: Pergamon.
- Muñoz, M. L., Gillam, R. B., Peña, E. D., & Gulley-Faehnle, A. (2003). Measures of language development in fictional narratives of Latino children. *Language, Speech, and Hearing Services in Schools*, 34(4), 332-342.
- Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Author. Retrieved from
https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf
- National Center for Education Statistics (2019). *The Nations Report Card: Reading 2019*. Washington D. C.: Institute of Education Sciences, U.S. Department of Education.
- National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington, DC: Authors.
- National Institute of Child Health and Human Development Early Child Care Research Network. (2005). Oral language and reading: Reply to Bracken (2005). *Developmental Psychology*, 41(6), 1000–1002.
- Newcomer, P., & Hammill, D. (1997). *Test of Language Development-Primary: 3*. Austin, TX: Pro-Ed.

- Nippold, M. A. (2014). *Language sampling with adolescents: Implications for intervention* (2nd ed.). San Diego, CA: Plural.
- Nippold, M. A., & Sun, L. (2010). Expository writing in children and adolescents: A classroom assessment tool. *Perspectives on Language Learning and Education*, 17(3), 100-107.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- Pesco, D., & Gagné, A. (2017). Scaffolding narrative skills: A meta-analysis of instruction in early childhood settings. *Early Education and Development*, 28(7), 773-793.
- Petersen, D. B. (2011). A systematic review of narrative-based language intervention with children who have language impairment. *Communication Disorders Quarterly*, 32(4), 207-220.
- Petersen, D. B., Gillam, S. L., & Gillam, R. B. (2008). Emerging procedures in narrative assessment: The index of narrative complexity. *Topics in Language Disorders*, 28(2), 115-130.
- Petersen, D. B., & Spencer, T. D. (2012). The narrative language measures: Tools for language screening, progress monitoring, and intervention planning. *Perspectives on Language Learning and Education*, 19(4), 119-129.
- Petersen, D. B., & Spencer, T. D. (2016). CUBED Assessment. Language Dynamics Group, LLC. <http://www.languagedynamicsgroup.com>
- Petersen, D. B., Spencer, T. D. (2019). *Narrative Language Measures Flow Chart (CUBED)*. Language Dynamics Group. <http://www.languagedynamicsgroup.com>
- Peterson, C. (1990). The who, when and where of early narratives. *Journal of Child Language*, 17(2), 433-455.
- Phillips Galloway, E., McClain, J. B., & Uccelli, P. (2020). Broadening the lens on the science of reading: A multifaceted perspective on the role of academic language in text understanding. *Reading Research Quarterly*, 55, S331-S345.
- Price, J. R., & Jackson, S. C. (2015). Procedures for obtaining and analyzing writing samples of school-age children and adolescents. *Language, Speech, and Hearing Services in Schools*, 46(4), 277-293.
- Proctor, C. P., Carlo, M., August, D., & Snow, C. (2005). Native Spanish-Speaking Children Reading in English: Toward a Model of Comprehension. *Journal of Educational Psychology*, 97(2), 246-256.
- Sanchez, S. V., Rodriguez, B. J., Soto-Huerta, M. E., Villarreal, F. C., Guerra, N. S., & Flores, B. B. (2013). A case for multidimensional bilingual assessment. *Language Assessment Quarterly*, 10(2), 160-177.

- Scheele, A. F., Leseman, P. P., Mayo, A. Y., & Elbers, E. (2012). The relation of home language and literacy to three-year-old children's emergent academic language in narrative and instruction genres. *The Elementary School Journal*, 112(3), 419-444.
- Schleppegrell, M. J. (2001). Linguistic features of the language of schooling. *Linguistics and education*, 12(4), 431-459.
- Schrank, F. A., & Wendling, B. J. (2018). The Woodcock–Johnson IV: Tests of cognitive abilities, tests of oral language, tests of achievement. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary Intellectual Assessment: Theories, Rests, and Issues* (pp. 383–451). The Guilford Press.
- Scott, C. M., & Balthazar, C. H. (2010). The grammar of information: Challenges for older students with language impairments. *Topics in Language Disorders*, 30(4), 288.
- Shanahan, T., Callison, K., Carriere, C., Duke, N. K., Pearson, P. D., Schatschneider, C., & Torgesen, J. (2010). Improving reading comprehension in Kindergarten through 3rd grade: IES practice guide. NCEE 2010-4038. *What Works Clearinghouse*.
- Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science*, 328(5977), 450-452.
- Snow, C. E., & Uccelli, P. (2008). The challenge of academic language. In Olson, D. R. & N. Torrance (Eds.), *The Cambridge Handbook of Literacy* (pp. 112-133). Cambridge: Cambridge University Press.
- Spencer, T. D., Kajian, M., Petersen, D. B., & Bilyk, N. (2013). Effects of an individualized narrative intervention on children's storytelling and comprehension skills. *Journal of Early Intervention*, 35(3), 243-269.
- Spencer, T. D., Petersen, D. B., & Adams, J. L. (2015). Tier 2 language intervention for diverse preschoolers: An early-stage randomized control group study following an analysis of response to intervention. *American Journal of Speech-Language Pathology*, 24(4), 619-636.
- Spencer, T. D., & Petersen, D. B. (2018). Bridging oral and written language: An oral narrative language intervention study with writing outcomes. *Language, Speech, and Hearing Services in Schools*, 49(3), 569-581.
- Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. *New directions in discourse processing*, 2(1979), 53-120.
- Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: evidence from a longitudinal structural model. *Developmental psychology*, 38(6), 934.
- Suggate, S. P. (2016). A meta-analysis of the long-term effects of phonemic awareness, phonics, fluency, and reading comprehension interventions. *Journal of Learning Disabilities*, 49(1), 77-96.

- Tavakol, M. & Dennicik, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, 2, 53-55.
- Tong, F., Lara-Alecio, R., Irby, B., Mathes, P., & Kwok, O. M. (2008). Accelerating early academic oral English development in transitional bilingual and structured English immersion programs. *American Educational Research Journal*, 45(4), 1011-1044.
- Truckenmiller, A. J., Park, J., Dabo, A., & Wu Newton, Y. C. (2019). Academic language instruction for students in grades 4 through 8: A literature synthesis. *Journal of Research on Educational Effectiveness*, 12(1), 135-159.
- Tucker, L. R., Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, 38, 1-10.
- Turner, M. (1996). *The literary mind: The origins of thought and language*. Oxford University Press.
- Uccelli, P., Barr, C. D., Dobbs, C. L., Galloway, E. P., Meneses, A., & Sánchez, E. (2014). Core academic language skills: An expanded operational construct and a novel instrument to chart school-relevant language proficiency in preadolescent and adolescent learners. *Applied Psycholinguistics*, 36(5), 1077-1109.
- Uccelli, P., Galloway, E. P., Barr, C. D., Meneses, A., & Dobbs, C. L. (2015). Beyond vocabulary: Exploring cross-disciplinary academic-language proficiency and its association with reading comprehension. *Reading Research Quarterly*, 50(3), 337-356.
- Uyanik, O., & Kandir, A. (2014). Adaptation of the Kaufman Survey of Early Academic and Language Skills to Turkish Children Aged 61 to 72 Months. *Educational Sciences: Theory and Practice*, 14(2), 682-692.
- Westby, C., Culatta, B., Lawrence, B., & Hall-Kenyon, K. (2010). Summarizing expository texts. *Topics in Language Disorders*, 30(4), 275-287.
- Westerveld, M. F., & Moran, C. A. (2013). Spoken expository discourse of children and adolescents: Retelling versus generation. *Clinical Linguistics & Phonetics*, 27(9), 720-734.
- Williams, J. P., Nubla-Kung, A. M., Pollini, S., Stafford, K. B., Garcia, A., & Snyder, A. E. (2007). Teaching cause—effect text structure through social studies content to at-risk second graders. *Journal of Learning Disabilities*, 40(2), 111-120.
- Woodcock, R. W. , & Johnson, M. B. (1989). Woodcock-Johnson Psycho-Educational Battery-Revised. Allen, TX: DLM.
- Woodcock, R. W., & Muñoz-Sandoval, A. F. (1993). *Woodcock-Muñoz Language Survey: Comprehensive manual*. Chicago, IL: Riverside Publishing Company.
- Zwiers, J. (2013). *Building academic language: Essential practices for content classrooms, grades 5-12*. John Wiley & Sons.

APPENDICES

Appendix A

NLM Flowchart (Front)

NLM Flow Chart

© 2020 Language Dynamics Group, LLC

Child Name/ID#: _____
 Grade: _____ Teacher: _____
 School: _____
 Examiner/Transcriber/Scorer: _____

	YEAR	MO	DAY
Date Tested			
Date of Birth			
Child's Age			

Sampling Context:
 Check all that apply

- ☐ Oral
☐ Written
☐ Personal Generation
☐ Fictional Generation
☐ Retell
☐ With pictures
☐ Without pictures
☐ Other: _____

Language Complexity Score: _____
 Narrative Structure Score: _____
 Punctuation Score: _____
 Capitalization Score: _____
 Average Word Rating: _____

TOTAL SCORE:
 (LC + NS)

WRITING CONVENTIONS (OPTIONAL)

PUNCTUATION	CAPITALIZATION	SPELLING
Number of words written _____	Number of words written _____	Sum of word ratings _____
Number of errors _____	Number of errors _____	Number of words written _____
PUNCTUATION SCORE Subtract number of errors from total number of words written.	CAPITALIZATION SCORE Subtract number of errors from total number of words written.	AVERAGE WORD RATING Divide the sum of word ratings by the total number of words written.
Calculate Errors Add up punctuation errors (up to 3 for each type) for when punctuation was needed and was used incorrectly (0 = not needed)	Calculate Errors Add up capitalization errors (up to 3 for each type)	Rating Scale Rate each word in the written sample using the rubric below
Period at end of sentence 0 1 2 3	Lowercase for regular word 1 2 3	0 Unconventional symbol. Contains vertical line, dot, circle instead of letter or number.
Question mark 0 1 2 3	Uppercase I 1 2 3	1 Conventional symbol. Contains at least one real letter or number, but is unrecognizable as a word. Examples: "4", "J", "1s", "B3n"
Apostrophe 0 1 2 3	Uppercase for first word of sentence 1 2 3	2 Phonetic representation. Contains one or more letters that are phonetically related to a recognizable word. Examples: "bb" or "bd" for bird, "r" for are
Quotation mark 0 1 2 3	Uppercase for proper names 1 2 3	3 Invented spelling. Contains two or more letters that represent most of the phonemes of a recognizable word. Must have a vowel and be easy to figure out. Examples: "bir" for bird, "gol" for girl
Comma in a list 0 1 2 3	Uppercase for holidays, days, and months 1 2 3	4 Conventional spelling. Spelled correctly.
	Uppercase for acronyms 1 2 3	

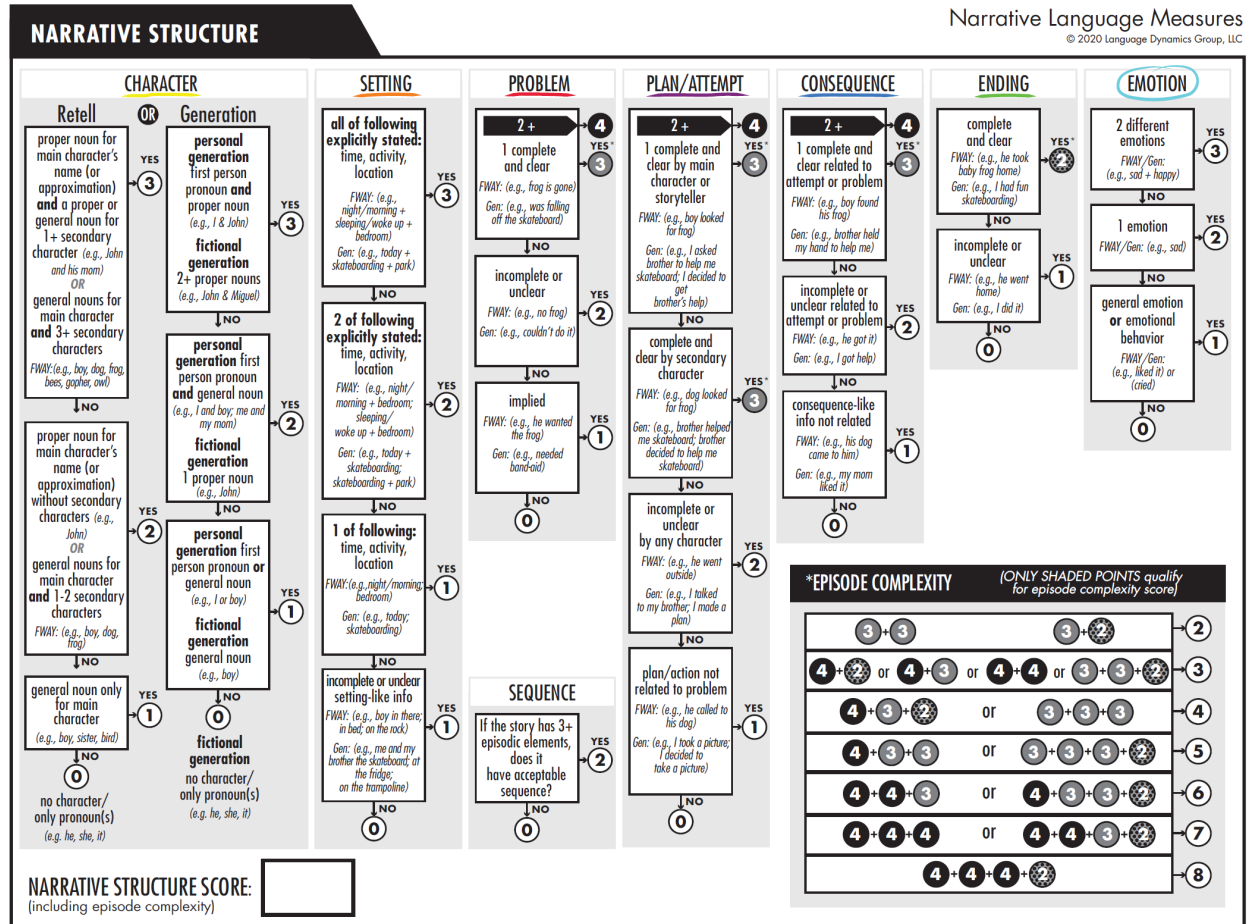
LANGUAGE COMPLEXITY

RELATIVE PRONOUNS	VERB/NOUN MODIFIERS	VOCABULARY/RHETORIC	TEMPORAL TIES	CAUSAL TIES	DIALOGUE
3+ instances of relative pronouns immediately after nouns (that, who, which, who's) (e.g., My friends who came to my house, are very nice.)	1+ instances of 2 consecutive descriptive modifiers (e.g., a big dirty dog) OR 2+ instances of single descriptive modifiers before a verb or a noun (e.g., We built our fort in the tall trees. We easily climbed the ladder.)	3+ less-common words/idioms/analogs/metaphors/similes (e.g., the owl swooped; her gaze was icy)	3+ instances of temporal words (that are often used in complex sentences) (when, after, before, while, as, until)	3+ instances of causal words (that are often used in complex sentences) (because, so (that), since, unless, although, even though)	2+ instances of dialogue; either 2 speakers or 2 separate instances of the same speaker
YES 3	YES 3	YES 3	YES 3	YES 3	YES 2
NO	NO	NO	NO	NO	NO
2 instances of relative pronouns immediately after nouns (that, who, which, who's) (e.g., We built a treehouse that has two rooms.)	1 instance of single descriptive modifier before a verb or a noun (e.g., a dirty dog) OR 2+ instances of single descriptive modifiers after a verb (e.g., We climbed the ladder easily)	2 less-common words/idioms/analogs/metaphors/similes (e.g., he tumbled; black as night)	2 instances of temporal words (that are often used in complex sentences) (when, after, before, while, as, until)	2 instances of causal words (that are often used in complex sentences) (because, so (that), since, unless, although, even though)	1 instance of dialogue
YES 2	YES 2	YES 2	YES 2	YES 2	YES 1
NO	NO	NO	NO	NO	NO
1 instance of relative pronoun immediately after noun (that, who, which, who's) (e.g., We built a fort, which was needed for a snowball battle.)	1 instance of single descriptive modifier after a verb (e.g., The dogs ran quickly.)	1 less-common words/idioms/analogs/metaphors/similes (e.g., he creak to the door)	1 instance of a temporal word (that is often used in complex sentences) (when, after, before, while, as, until)	1 instance of a causal word (that is often used in complex sentences) (because, so (that), since, unless, although, even though)	
YES 1	YES 1	YES 1	YES 1	YES 1	
NO	NO	NO	NO	NO	
0	0	0	0	0	

LANGUAGE COMPLEXITY SCORE: _____

Appendix B

NLM Flowchart (Back)



Appendix C

ELM Flowchart (Front)

ELM Flow Chart

© 2020 Language Dynamics Group, LLC

Child Name/ID#: _____
 Grade: _____ Teacher: _____
 School: _____
 Examiner/Transcriber/Scorer: _____

	YEAR	MO	DAY
Date Tested			
Date of Birth			
Child's Age			

Sampling Context:
 Check all that apply

- ☐ Oral
☐ Written
☐ Generation
☐ Retell
☐ With pictures
☐ Without pictures
☐ Other: _____

Language Complexity Score: _____

Passage Structure Score: _____

Punctuation Score: _____

Capitalization Score: _____

Average Word Rating: _____

TOTAL SCORE:
 (LC + PS)

LANGUAGE COMPLEXITY

RELATIVE PRONOUNS	VERB/NOUN MODIFIERS	VOCABULARY	TEMPORAL TIES	CAUSAL TIES	TRANSITIONS
<p>3 + instances of relative pronouns immediately after nouns (that, who, which, who's) (e.g., <i>Monkeys, who are agile creatures, swing in the trees.</i>)</p> <p>YES 3</p> <p>NO</p> <p>2 instances of relative pronouns immediately after nouns (that, who, which, who's) (e.g., <i>Many snakes live in trees that have enough leaves to disguise them.</i>)</p> <p>YES 2</p> <p>NO</p> <p>1 instance of relative pronoun immediately after a noun (that, who, which, who's) (e.g., <i>Early humans began building cities, which allowed them to stay in one place.</i>)</p> <p>YES 1</p> <p>NO 0</p>	<p>1+ instances of 2 consecutive descriptive modifiers (e.g., <i>a large, dirty monkey</i>) OR</p> <p>2+ instances of single descriptive modifiers before a verb or a noun (e.g., <i>Monkeys are found in the tall trees. We easily swing from limb to limb.</i>)</p> <p>YES 3</p> <p>NO</p> <p>1 instance of single descriptive modifier before a verb or a noun (e.g., <i>The large monkeys</i>) OR</p> <p>2+ instances of single descriptive modifiers after a verb (e.g., <i>Monkeys can climb quickly.</i>)</p> <p>YES 2</p> <p>NO</p> <p>1 instance of single descriptive modifier after a verb (e.g., <i>Monkeys swing easily in the trees.</i>)</p> <p>YES 1</p> <p>NO 0</p>	<p>3 + less-common domain-specific words related to the topic (e.g., <i>their natural habitat; they change their form during metamorphosis.</i>)</p> <p>YES 3</p> <p>NO</p> <p>2 less-common domain-specific words related to the topic (e.g., <i>high body temperature; they grow crops.</i>)</p> <p>YES 2</p> <p>NO</p> <p>1 less-common domain-specific word related to the topic (e.g., <i>trash goes to landfills.</i>)</p> <p>YES 1</p> <p>NO 0</p>	<p>3 + instances of temporal words (that are often used in complex sentences) (when, after, before, while, as, until)</p> <p>YES 3</p> <p>NO</p> <p>2 instances of temporal words (that are often used in complex sentences) (when, after, before, while, as, until)</p> <p>YES 2</p> <p>NO</p> <p>1 instance of a temporal word (that is often used in complex sentences) (when, after, before, while, as, until)</p> <p>YES 1</p> <p>NO 0</p>	<p>3 + instances of causal words (that are often used in complex sentences) (because, so (that), since, unless, although, even though)</p> <p>YES 3</p> <p>NO</p> <p>2 instances of causal words (that are often used in complex sentences) (because, so (that), since, unless, although, even though)</p> <p>YES 2</p> <p>NO</p> <p>1 instance of a causal word (that is often used in complex sentences) (because, so (that), since, unless, although, even though)</p> <p>YES 1</p> <p>NO 0</p>	<p>2 + instances of transition words/phrases (e.g., <i>therefore, similarly, as a result, however, for example, likewise, in contrast</i>)</p> <p>YES 4</p> <p>NO</p> <p>1 instance of a transition word/phrase (e.g., <i>therefore, similarly, as a result, however, for example, likewise, in contrast</i>)</p> <p>YES 2</p> <p>NO 0</p>

LANGUAGE COMPLEXITY SCORE: _____

WRITING CONVENTIONS (OPTIONAL)

PUNCTUATION	CAPITALIZATION	SPELLING
<p>Number of words written _____</p> <p>Number of errors _____</p> <p>PUNCTUATION SCORE Subtract number of errors from total number of words written.</p> <p>Calculate Errors Add up punctuation errors (up to 3 for each type) for when punctuation was needed and was used incorrectly (0 = not needed)</p> <p>Period at end of sentence 0 1 2 3</p> <p>Question mark 0 1 2 3</p> <p>Apostrophe 0 1 2 3</p> <p>Quotation mark 0 1 2 3</p> <p>Comma in a list 0 1 2 3</p>	<p>Number of words written _____</p> <p>Number of errors _____</p> <p>CAPITALIZATION SCORE Subtract number of errors from total number of words written.</p> <p>Calculate Errors Add up capitalization errors (up to 3 for each type)</p> <p>Lowercase for regular word 1 2 3</p> <p>Uppercase I 1 2 3</p> <p>Uppercase for first word of sentence 1 2 3</p> <p>Uppercase for proper names 1 2 3</p> <p>Uppercase for holidays, days, and months 1 2 3</p> <p>Uppercase for acronyms 1 2 3</p>	<p>Sum of word ratings _____</p> <p>Number of words written _____</p> <p>AVERAGE WORD RATING Divide the sum of word ratings by the total number of words written.</p> <p>Rating Scale Rate each word in the written sample using the rubric below</p> <p>0 Unconventional symbol. Contains vertical line, dot, circle instead of letter or number.</p> <p>1 Conventional symbol. Contains at least one real letter or number, but is unrecognizable as a word. Examples: "4", "j", "ls", "8Sn"</p> <p>2 Phonetic representation. Contains one or more letters that are phonetically related to a recognizable word. Examples: "bb" or "bd" for bird, "i" for are</p> <p>3 Invented spelling. Contains two or more letters that represent most of the phonemes of a recognizable word. "Must have a vowel and be easy to figure out. Examples: "hir" for bird, "gol" for girl</p> <p>4 Conventional spelling. Spelled correctly.</p>

Appendix D

ELM Flowchart (Back)

PASSAGE STRUCTURE




Expository Language Measures
© 2020 Language Dynamics Group, LLC

MAIN IDEA	INFORMATION UNITS	DEFINITIONS & EXAMPLES	PASSAGE COHESION	CONCLUDING STATEMENT
<div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> 2 + complete and clear main ideas directly related to the pictures/topic or from model passage <i>(e.g., Caterpillars turn into butterflies during metamorphosis.)</i> </div> <div style="text-align: right; font-weight: bold;">YES 3</div> <div style="text-align: center; margin: 5px 0;">↓ NO</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> 1 complete and clear main idea directly related to the pictures/topic or from model passage <i>(e.g., Humans have five senses to help them learn about the world.)</i> </div> <div style="text-align: right; font-weight: bold;">YES 2</div> <div style="text-align: center; margin: 5px 0;">↓ NO</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> 1 + incomplete or unclear main idea <i>(e.g., Tigers live in the jungle.)</i> </div> <div style="text-align: right; font-weight: bold;">YES 1</div> <div style="text-align: center; margin: 5px 0;">↓ NO</div> <div style="text-align: center; font-weight: bold; font-size: 1.2em;">0</div>	<p style="text-align: center; font-weight: bold;">INFORMATION UNITS</p> <p style="text-align: center; font-size: 0.8em;">Refer to the Scoring Manual for detailed instructions for scoring information units.</p> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> Factual Unit = a clause, containing a subject and a verb, that conveys one piece of factual information or is presented like it is factual, whether or not it is accurate. <i>e.g., Tigers are carnivores (1). You want to stay away from them (2). When tigers stroll through the jungle (3), they carefully search for their next meal (4).</i> </div> <div style="text-align: center; font-weight: bold; font-size: 0.8em;"> 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 </div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> Picture Description Unit = a clause, containing a subject and a verb, that explicitly describes what is shown in the picture(s) or directly references the picture. <i>e.g., In this picture, it looks like they dug the ground (1). Here the mushroom grows (2). This one is healthy (3).</i> </div> <div style="text-align: center; font-weight: bold; font-size: 0.8em;"> 1 2 3 4 5 6 7 8 9 10 </div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> Narrative Unit = a clause, containing a subject and a verb, that tells about a specific real or imaginary event in past tense; the subject is often a character. <i>e.g., My grandpa and I went camping (1). There was a lake by our campsite (2). They decided to go get help (3).</i> </div> <div style="text-align: center; font-weight: bold; font-size: 0.8em;"> 1 2 3 4 5 6 7 8 9 10 </div> <div style="text-align: center; margin-top: 10px;"> Number of Information Units ÷ 2 = </div>	<p style="text-align: center; font-weight: bold;">DEFINITIONS & EXAMPLES</p> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> Use of at least 1 definition AND Use of at least 1 example <i>(e.g., Flora refers to the plants that live in an area. In desert climates, the flora includes a variety of cacti.)</i> </div> <div style="text-align: right; font-weight: bold;">YES 3</div> <div style="text-align: center; margin: 5px 0;">↓ NO</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> Use of a definition <i>(e.g., Flora refers to the plants that live in an area.)</i> </div> <div style="text-align: right; font-weight: bold;">YES 2</div> <div style="text-align: center; margin: 5px 0;">OR</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> Use of an example <i>(e.g., Monkeys and gorillas are some of the fauna living in jungles.)</i> </div> <div style="text-align: right; font-weight: bold;">YES 2</div> <div style="text-align: center; margin: 5px 0;">↓ NO</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> Use of an incomplete or unclear definition or example <i>(e.g., It means the animals.)</i> </div> <div style="text-align: right; font-weight: bold;">YES 1</div> <div style="text-align: center; margin: 5px 0;">↓ NO</div> <div style="text-align: center; font-weight: bold; font-size: 1.2em;">0</div>	<p style="text-align: center; font-weight: bold;">PASSAGE COHESION</p> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> A main idea is stated and all information units support the main idea </div> <div style="text-align: right; font-weight: bold;">YES 3</div> <div style="text-align: center; margin: 5px 0;">↓ NO</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> There is no main idea but most or all of the information units are about the same topic OR A main idea is stated and some of the information supports it </div> <div style="text-align: right; font-weight: bold;">YES 2</div> <div style="text-align: center; margin: 5px 0;">↓ NO</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> There is no main idea and only some of the information units are about the same topic </div> <div style="text-align: right; font-weight: bold;">YES 1</div> <div style="text-align: center; margin: 5px 0;">↓ NO</div> <div style="text-align: center; font-weight: bold; font-size: 1.2em;">0</div>	<p style="text-align: center; font-weight: bold;">CONCLUDING STATEMENT</p> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;"> Does the passage have a concluding statement? </div> <div style="text-align: right; font-weight: bold;">YES 1</div> <div style="text-align: center; margin: 5px 0;">↓ NO</div> <div style="text-align: center; font-weight: bold; font-size: 1.2em;">0</div>
<div style="text-align: right; font-weight: bold; font-size: 1.2em;">PASSAGE STRUCTURE SCORE: </div>				

EXPOSITION TYPE
 What type of exposition best fits this sample?
☐ How To
☐ Description
☐ Sequence
☐ Comparison
☐ Cause/Effect
☐ Problem/Solution
Does not count toward Passage Structure Score

Appendix E

Narrative Elicitation Script Example

SCRIPT FOR NARRATIVE RETELL		Oral 
<input type="checkbox"/> Switch the digital voice recorder on. <input type="checkbox"/> Speak into the recorder the child's ID#, the date, your examiner ID# and the task (i.e., Narrative-Retell-Oral).		
<div> <div> <hr/> NARRATIVE RETELL 1 <hr/> </div> <div> <hr/> NARRATIVE RETELL 2 <hr/> </div> </div>		
<input type="checkbox"/> Place 3 picture sets in front of the student on the table in the following array. 	<input type="checkbox"/> Place the 2 remaining picture sets on the table. 	
<input type="checkbox"/> Say, "We are going to tell a story about the pictures on one of these cards. Which pictures would you like us to tell a story about?"	<input type="checkbox"/> Say, "We are going to tell a story about the pictures on one more card. Which pictures would you like us to tell a story about?"	
<input type="checkbox"/> Student chooses a picture set.	<input type="checkbox"/> Student chooses a picture set.	
<input type="checkbox"/> Clear the 2 unchosen picture sets from the table and say, "You chose Picture Set (A, B, C...)," correctly identifying the letter in the lower right corner of the chosen picture set.	<input type="checkbox"/> Clear the unchosen picture set from the table and say, "You chose Picture Set (A, B, C...)," correctly identifying the letter in the lower right corner of the chosen picture set.	
<input type="checkbox"/> Place the chosen picture set in front of the student.	<input type="checkbox"/> Place the chosen picture set in front of the student.	
<input type="checkbox"/> Say, "I'm going to tell you a story about these pictures. Please listen carefully. When I'm done, you are going to tell me the same story. Are you ready?"	<input type="checkbox"/> Say, "I'm going to tell you a story about these pictures. Please listen carefully. When I'm done, you are going to tell me the same story. Are you ready?"	
<input type="checkbox"/> Read the story word for word at a moderate pace with normal inflection.	<input type="checkbox"/> Read the story word for word at a moderate pace with normal inflection.	
<input type="checkbox"/> Say, "Thanks for listening. Now you tell me the same story. There is no right or wrong answer." <ul style="list-style-type: none"> - Use only neutral prompts (e.g., "uh huh") while the student retells the story. - If the student needs encouragement, say, "There is no right or wrong answer. Just do the best you can." 	<input type="checkbox"/> Say, "Thanks for listening. Now you tell me the same story. There is no right or wrong answer." <ul style="list-style-type: none"> - Use only neutral prompts (e.g., "uh huh") while the student tells his/her story. - If the student needs encouragement, say, "There is no right or wrong answer. Just do the best you can." 	
<input type="checkbox"/> When the student stops talking for 5-7 seconds, ask, "Can you tell me anything else?" <ul style="list-style-type: none"> - If the student indicates "yes" or continues talking, allow him/her to finish. 	<input type="checkbox"/> When the student stops talking for 5-7 seconds, ask, "Can you tell me anything else?" <ul style="list-style-type: none"> - If the student indicates "yes" or continues talking, allow him/her to finish. 	
<input type="checkbox"/> If the student says "no" or when the student stops talking for 5-7 seconds a second time, ask, "Are you finished?" <ul style="list-style-type: none"> - If the student indicates "no," allow him/her to finish. - If the student indicates "yes," remove the picture set from the table. 	<input type="checkbox"/> If the student says "no" or when the student stops talking for 5-7 seconds a second time, ask, "Are you finished?" <ul style="list-style-type: none"> - If the student indicates "no," allow him/her to finish. 	
<input type="checkbox"/> If the student indicates "yes" or when the student is finished, stop the recorder and say, "Thanks for retelling the stories." <ul style="list-style-type: none"> - Remove the picture set from the table. 		
<input type="checkbox"/> Write the audio file #, chosen picture sets, date, and your initials in the appropriate spaces on the tracking sheet.		



Appendix F

Expository Elicitation Script Example

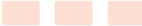
SCRIPT FOR

EXPOSITORY GENERATION


Oral 

- ☐ Switch the digital voice recorder on.
- ☐ Speak into the recorder the child's ID#, the date, your examiner ID#, and the task (i.e., Expository-Generation-Oral).

EXPOSITORY GENERATION 1

- ☐ Place 3 picture sets in front of the student on the table in the following array. 
- ☐ Say, "You are going to talk about the pictures on one of these cards. Which pictures would you like to talk about?"
- ☐ Student chooses a picture set.
- ☐ Clear the 2 unchosen picture sets from the table and say, "You chose Picture Set (A, B, C...)," correctly identifying the letter in the lower right corner of the chosen picture card.
- ☐ Place the chosen picture set in front of the student.
- ☐ Inserting the main idea for the chosen picture set, say, "The main idea of this set of pictures is _____. You are going to tell me everything you can about _____. I'm going to give you some time to think about it, and then I'll ask you to tell me. There is no right or wrong answer. Do you have any questions?"
- ☐ Set the timer for 30 seconds and start it.
 - If the student asks to begin before 30 seconds have passed, say, "Are you ready to start?" If the student indicates "yes", skip to, "Okay. Please tell me everything you can about _____."
- ☐ After 30 seconds, say, "Okay. Please tell me everything you can about _____."
 - Use only neutral prompts (e.g., "uh huh") while the student tells what he/she knows about the pictures.
 - If the student needs encouragement, say, "There is no right or wrong answer. Just do the best you can."
- ☐ When the student stops talking for 5-7 seconds, ask, "Can you tell me anything else?"
 - If the student indicates "yes" or continues talking, allow him/her to finish.
- ☐ If the student says "no" or when the student stops talking for 5-7 seconds a second time, ask, "Are you finished?"
 - If the student indicates "no," allow him/her to finish.
 - If the student indicates "yes," remove the picture set from the table.

EXPOSITORY GENERATION 2

- ☐ Place the 2 remaining picture sets on the table. 
- ☐ Say, "You are going to talk about the pictures on one more card. Which pictures would you like to talk about this time?"
- ☐ Student chooses a picture set.
- ☐ Clear the unchosen picture set from the table and say, "You chose Picture Set (A, B, C...)," correctly identifying the letter in the lower right corner of the chosen picture card.
- ☐ Place the chosen picture set in front of the student.
- ☐ Inserting the main idea for the chosen picture set, say, "The main idea of this set of pictures is _____. You are going to tell me everything you can about _____. I'm going to give you some time to think about it, and then I'll ask you to tell me. There is no right or wrong answer. Do you have any questions?"
- ☐ Start the timer.
 - If the student asks to begin before 30 seconds have passed, say, "Are you ready to start?" If the student indicates "yes", skip to, "Okay. Please tell me everything you can about _____."
- ☐ After 30 seconds, say, "Okay. Please tell me everything you can about _____."
 - Use only neutral prompts (e.g., "uh huh") while the student tells what he/she knows about the pictures.
 - If the student needs encouragement, say, "There is no right or wrong answer. Just do the best you can."
- ☐ When the student stops talking for 5-7 seconds, ask, "Can you tell me anything else?"
 - If the student indicates "yes" or continues talking, allow him/her to finish.
- ☐ If the student says "no" or when the student stops talking for 5-7 seconds a second time, ask, "Are you finished?"
 - If the student indicates "no," allow him/her to finish.
- ☐ If the student indicates "yes" or when the student is finished, stop the recorder and say, "Thanks for talking about the pictures."
 - Remove the picture set from the table.

- ☐ Write the audio file #, chosen picture sets, date, and your initials in the appropriate spaces on the tracking sheet.

