The Dissertation Committee for Laura A. McFarland Certifies that this is the approved version of the following dissertation:

Tell Me a Story: Scoring and Analysis of the English Oral Narrative Skills of Second Grade Spanish-Speaking English-Language Learners

Committee:

Sylvia F. Thompson, Supervisor

Alba A. Ortiz, Co-Supervisor

Andrea L. Flower

Amanda L. Little

Phyllis M. Robertson

Cheryl Y. Wilkinson

Anita M. Perez

Tell Me a Story: Scoring and Analysis of the English Oral Narrative Skills of Second Grade Spanish-Speaking English-Language Learners

by

Laura Ann McFarland, B.A., M.Ed.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2013

Dedication

To Fabrizio, Juan Pablo, and my other first grade students in Lima, Peru, who taught me to see, hear and to seek to understand the many faces and languages of genius. And, to Krissy, Mad Dog, Lacey, and Spud – you make a truly happy home.

Acknowledgements

There are several individuals to whom I am deeply indebted for the support, guidance, and friendship they have provided throughout this endeavor. First and foremost, I wish to thank my mentor and advisor, Dr. Alba Ortiz. Thank you, Dr. Ortiz, for entrusting me with your data and for mentoring me from the very inception of this study. I could not have accomplished this without you. Dr. Phyllis Robertson, you are truly an angel, who has always had my back. Thank you for being my friend and advocate and excellent colleague all these years. I would like to thank the rest of my committee as well, especially Dr. Sylvia Thompson, for her supervision and her guidance, and Dr. Cheryl Wilkinson, for her always thoughtful and careful consideration and helpful feedback. Drs. Andrea Flower, Amanda Little, and Anita Perez – I thank each of you for participating on my committee and for your excellent questions and comments, which have made this a better study. I would also like to thank Dr. Shernaz Garcia and the rest of the Multicultural Special Education community, students and faculty alike, for 'walking the talk'. You are an inspiration.

On a personal note, I thank my wonderful parents for always believing in me and setting the bar high. I thank my brothers and their families for keeping me sane and giving me perspective. And I thank my dear friends, especially Natalie, Dominique, Kim, Gisela, Forrest, Rhonda, Cindy, Lisa, and, last but not least, Krissy. You bring me great joy. Ultimately, I give thanks to God, the Creator and greatest narrator of all – by whom, in whom, and with whom we live, breathe, and have our being, and who establishes the work of our hands.

Tell Me a Story: Scoring and Analysis of the English Oral Narrative Skills of Spanish-Speaking English-Language Learners

Laura A. McFarland, Ph.D.

The University of Texas at Austin, 2013

Supervisor: Sylvia F. Thompson

Co-Supervisor: Alba A. Ortiz

Competence with oral narrative discourse is associated with both reading comprehension and academic achievement in general. However, most research on narratives has been conducted with monolingual English speaking children and the theoretical frameworks used to measure narrative skills are predominantly based on what is known about the narrative skills of this population. There has been much less research examining the narrative skills of English language learners (ELLs) and how to best assess these skills. This exploratory study examined the characteristics of the English oral narratives of Spanish-speaking ELLs (SS-ELLs). The narrative data are a subset of data collected as part of a model demonstration project conducted by faculty from The University of Texas in partnership with a central Texas school district. The student sample included 42 SS-ELLs enrolled in a bilingual second grade classroom. Transcripts of stories told in response to a picture prompt were coded and analyzed according to three narrative scoring systems: story grammar analysis, Narrative Assessment Profile, and Narrative Scoring Scheme. Results of these analyses were used to: 1) describe the qualities of the English oral stories of Spanish-speaking ELLs in terms of their organization and production; 2) examine how each scoring system characterizes the

sample in terms of expected performance according to its criteria; 3) identify the stable features of narratives whose performance is rated consistently across measures and aspects of scoring systems that are well matched and mismatched to evaluate those features; and 4) identify characteristics of scoring systems that produce information that is useful to instructional planning for SS-ELLs in ESL settings. Recommendations for analyzing the oral narratives of SS-ELLs in ways that are reliable and useful to instructional planning are offered.

Table of Contents

CHAPTER ONE **1**

Introduction 1

Narrative Assessment 4

Elicitation and Analysis of Narratives 5

Language Productivity and Organization (Microstructural) Skills 6

Narrative Organization (Macrostructural) Skills 7

Potentials and Limitations of Existing Methods of Narrative Assessment 8

Statement of the Problem 10

Research Questions 12

Significance of the Study 13

Summary 14

CHAPTER TWO **16**

Literature Review 16

The Growing Population of Spanish-Speaking English Language Learners 17

School Related Outcomes 18

Pervasive Achievement Gap 18

Disproportionate Representation in Special Education 23

Factors Contributing to Inappropriate Special Education Outcomes 29

Distinguishing Reading Difficulties Arising from Limited English Proficiency from Those Arising from Reading Disabilities in Spanish-Speaking English Language Learners 29

Culturally and Linguistically Responsive Assessment Practices for ELLs 33

Considerations in Conducting Narrative Assessments with ELLs 34

The Need to Investigate the Narratives of ELLs across the School Years and in Relation to Reading 35

Cross-Linguistic Relationships between Narrative Performance and Reading Performance of ELLs 38

Gaps in the Research on the Narrative Skills of ELLs 39
Insufficient Descriptions of the Samples 40
Lack of Studies Investigating the Narrative Skills of ELLs with and without Learning Disabilities and the Relationship of those skills to Reading.....41
Incomparability of Findings between Studies due to Important Methodological Differences.....42
The Assessment of Children's Oral Narratives 44
Scoring Systems Used in the Present Study 45
Story Grammar Analysis.....45
Narrative Assessment Profile.....54
Narrative Scoring Scheme 60
Characteristics of the Narratives of Spanish-Speaking English Language Learners.....63
Description of Studies Included in the Review.....65
Language Productivity Characteristics of ELLs' Narratives 66
Story Grammar Characteristics of SS-ELLs' Narratives.....67
Summary 71
CHAPTER THREE 72
Method 72
Research Questions 72
Context for the Study 73
Model Demonstration Project 73
Participating Sites 74
Tell me a Story: Scoring and Analysis of English Oral Narrative Skills of Second Grade Spanish-Speaking English Language Learners 74
Research Approval 74
Participants.....75
Language Proficiency Level of Participants 75
Data Sources 76
Instruments.....77

Confidentiality of Data 78
Narrative Analysis 78
Preparation of Story Transcripts 78
Narrative Scoring 80
Story Grammar Analysis 81
Narrative Assessment Profile 83
Narrative Scoring Scheme 85
Reliability 86
Selection of a Co-Rater 86
Training Using Transcripts Generated by Select Non-Participant Sample 87
Selection of the Reliability Sample 88
Process of Establishing Reliability 88
Transcript Segmentation 88
Story Grammar Analysis 90
Narrative Assessment Profile 95
Narrative Scoring Scheme 99
Measuring Inter-Rater Reliability 106
Transcript Segmentation 107
Story Grammar Analysis 107
Narrative Assessment Profile 109
Narrative Scoring Scheme 110
Data Analysis 111
Data Preparation 111
Descriptive Statistics 111
Answering Research Questions 111
Summary of Method 113

CHAPTER FOUR 115

Results 115
The Characteristics of Spanish-Speaking English Learners' Narratives 118

General Findings....118
Microstructural Characteristics of SS-ELLs' Narratives....118
Macrostructure Characteristics of Narratives....124
Characteristics of SS-ELLs' Oral English Narratives as Measured by Story Grammar Analysis....127
Characteristics of SS-ELLs' Oral English Narratives as Measured by the Narrative Assessment Profile (NAP)....134
Characteristics of SS-ELLs' Oral English Narratives as Measured by the Narrative Scoring Scheme (NSS)....145
Stable Features Across Measures....166
Stratifying the Sample....166
Cases Rated Consistently Average, Below Average, or Above Average across Scoring Systems....170
Microstructural Features of Low, Average, and Above Average Oral Narrative Performance....172
Macrostructural Features of Low, Average, and Above Average Oral Narrative Performance....176
Summary....184

CHAPTER FIVE....187

Discussion....187
Summary of the Study....187
The English Oral Narrative Performance of Elementary Age Spanish-Speaking English Language Learners....189
The Narrative Scoring Scheme as a Measure of the Oral Narrative Skills of Elementary Age Spanish-Speaking English Language Learners....196
Incongruence between Stable Features of SS-ELLs' Oral English Narratives and Properties of Narrative Scoring Systems....202
Criteria for a High Quality Instrument for the Assessment of the English Oral Narrative Performance of SS-ELLs....208
Limitations of the Study and Recommendations for Future Research....210
Summary....214
Concluding Remarks....215

Appendix A: Picture Prompts 216

Appendix B: Scoring Rubrics 217

Appendix B1: Story Grammar Analysis Modified Decision Guide and Rubric 217

Appendix B2: Narrative Assessment Profile Modified Coding Criteria 223

Appendix B3: Narrative Scoring Scheme Modified Rubric 226

Appendix C: Transcript Coding Decision Rules 231

References 236

CHAPTER ONE

Introduction

Experts in the language and literacy development of English language learners (ELLs) have recommended that a variety of formal and informal assessments, in both the child's native language and in English, be used when considering whether reading difficulties arise from an underlying reading-related learning disability or whether they are artifacts of the language acquisition process itself (August & Shanahan, 2006; Cummins, 1979; Figueroa, 2002; Ortiz, 1997; Ortiz & Yates, 2001; Pray, 2005; Stockman, 1996). Assessment of storytelling skills has been recommended as a particularly valuable informal language measure because it provides a sample, not only of surface language abilities such as grammar, syntax, and vocabulary, but also of macrostructural features related to children's ability to organize information episodically through extended text, without the help of a dialogic partner (Hughes, McGillivray, & Schmidek, 1997; McCabe & Bliss, 2003; McCabe & Rollins, 1994; Ortiz, 1997; Ortiz & Yates, 2002; Peterson, Gillam, & Gillam, 2008). In other words, when children are asked to tell a story, they must organize events and provide sufficient detail for the listener to make sense of the story with little need for clarification. Monologic in nature, storytelling requires skills very similar to those needed for reading comprehension (Damico, 1991). For example, attention must be given to reference and cohesion, agreement, and the organization of content over extended speech. All necessary information must be present in the delivery of the story. This requires the narrator to plan and organize information cognitively before delivering it so that it is comprehensible to the listener. Because skilled narrators and those with proficiency in reading

comprehension draw on similar capabilities, oral narrative skills are considered to be closely related to reading (Boudreau, 2008; Owens, 2010; Stockman, 1996; Westby, 1992).

Storytelling is a naturalistic task and stories provide a wealth of information about children's oral language skills, including those metalinguistic skills related to formal literacy. Narrative tasks are therefore particularly useful in the informal assessment of ELLs who are struggling to learn, read, and write in English (Figueroa, 2002; Ortiz, 1997; Ortiz & Yates, 2002; Pray, 2005; Stockman, 1996). Narrative samples provide information about ELLs' communicative competency as well as their cognitive academic language proficiency (CALP), the latter of which is specifically related to literacy and school achievement (Cummins, 1979; Cummins, 1984; Ortiz, 1997). While numerous studies have described the narrative skills of monolingual, English speaking students and have explored the relationships of their narrative skills to oral language proficiency, literacy, and to academic success, very few studies have documented the English oral narrative skills of English language learners.

This is an important gap to fill. Given the increasing prevalence of ELLs attending the nation's public schools, improving academic outcomes for this population has become a national priority. Indeed one of the most pressing challenges facing our public education system is that of teaching a rapidly growing population of language-minority students to read and write in English with such proficiency that they can participate fully in school (August & Hakuta, 1997; August & Shanahan, 2006). Public schools have seen a dramatic increase in the number of ELLs, or students with limited proficiency in English. These are children who speak a language other than English at home and who speak English with difficulty (Aud, Hussar, Kena, Bianco, Frohlich, et al., 2011). As of 2009, there were 11.2 million children, ages 5-17, who spoke a language

other than English at home and nearly 2.7 million of those spoke English with difficulty (U.S. Department of Commerce Census Bureau, 2009). Over 4.6 million (K-12) students participated in programs for English language learners during the 2010-2011 school year, representing nearly 10% of the total public school student population (U.S. Department of Education, 2012). While more than 400 languages are spoken by ELLs nationwide (Kindler, 2002), an estimated 73% are Spanish-speaking children (Aud et al., 2011).

Spanish-speaking ELLs (SS-ELLs) constitute the largest language-minority group in the nation's public schools. They fare particularly poorly when it comes to literacy, academic achievement, and graduation rates (Aud et al., 2011; Snyder & Dillow, 2011). Furthermore, there is concern that they are either underserved or inappropriately served in programs such as special education (Artiles, Kozleski, Trent, Osher, & Ortiz, 2010). The challenge of educating SS-ELLs issues, in part, from an incomplete understanding of the ways in which English language acquisition, especially the acquisition of literacy-related oral language skills, interacts with English literacy development (Klingner, Hoover, & Baca, 2008). This fragmented understanding makes it difficult to interpret patterns of performance and is compounded by limited means of assessing performance patterns over time. Culturally and linguistically appropriate assessment practices and the development of a robust empirical knowledge base about this population's language and literacy characteristics are thus key to addressing this challenge (Artiles & Ortiz, 2002; Harry & Klingner, 2006; Ortiz, 1997). Routine classroom assessment of SS-ELLs' oral narratives accompanied by a better empirical understanding of the characteristics of their typical or expected performance on such measures will likely contribute to more accurate interpretations of oral language and literacy performance at school, resulting in improved services and instructional programming for this population, which will, in turn, promote better achievement outcomes.

The current study seeks to contribute to this end by applying three methods of narrative assessment to a set of stories told by second grade SS-ELLs in response to a picture prompt. The stories, which were collected as part of a prior study, provide a diverse sample of SS-ELLs storytelling abilities in English. Asked to tell a story about a static picture, children's resulting narratives range from simple descriptions of objects and actions in the picture, to well developed stories integrating the various depicted characters and actions into meaningful episodes with problems, goal-driven actions, resolutions, and inferences. The ability to integrate the various components of a static picture in such a manner separates good narrators from poor ones. Although communicative competence in English may facilitate the production of better and more complete narratives, it in no way guarantees it. Likewise, even children whose communicative competence in English is relatively low may produce narratives in that language that are well structured, organized, and that convey a coherent and meaningful story. In this way, the assessment of SS-ELLs' English narratives can help us see beyond those surface language deficits that occur as a natural stage of the second language acquisition process and, instead, observe the cognitive and metalinguistic resources SS-ELLs bring to the storytelling task despite the limitations imposed by their actual levels of proficiency with English syntax, vocabulary, and pragmatic conventions.

NARRATIVE ASSESSMENT

Narrative discourse is a type of extended monologue, as opposed to an interactive dialogue, that school children are expected to comprehend and produce (Owens, 2010; Ortiz, 1997; Westby, 1984, 1992). The production of narratives requires more than surface language skills such as phonology, morphology, and syntax; it requires the ability to organize and maintain an extended discourse without the support of a conversational

partner (Westby, 1992). The discourse skills required to produce oral narratives are thus considered to be those oral language skills that are most closely related to literacy (Ortiz, 1997; Owens, 2010; Westby, 1984, 1992). Narrative assessments are recommended for the following reasons: (a) narrative skill is associated with other academic skills; (b) narratives occur naturally both within and out of school settings and thus have ecological validity; (c) narrative production is a rigorous test of various levels and aspects of language form, content, and use; (d) the difficulty of narrative tasks and the levels of support provided to help children produce narratives can be adjusted to reveal the optimal degree of support needed; and (e) narrative comprehension and production can both be assessed so that relative strengths related to receptive and expressive language can be compared (Hughes et al., 1997).

Elicitation and Analysis of Narratives

Narrative samples provide data from which several features of oral language can be measured (Miller et al., 2006), including but not limited to narrative discourse skills, syntax and vocabulary (Dickinson & McCabe, 2001). Narrative samples are often elicited by asking children to recount a personal event (McCabe & Bliss, 2003) or to retell or generate a fictional story. Pictures or other prompts (e.g., story starters) are sometimes used to stimulate ideas (Ely, Wolf, McCabe, & Melzi, 2000; Hughes et al., 1997). Once narrative samples are collected, they are transcribed and analyzed according to which features of the narratives are of interest to the assessor.

There are two general approaches to the analysis of narrative transcripts: those that examine discrete language skills associated with narrative microstructure (e.g., syntax, referential cohesion, number of words and number of different words produced, subordination, and grammaticality); and those that examine narrative macrostructure,

including the organization of episodes and the inclusion of essential story grammar elements. Analyses of narrative macrostructure involve what Heilmann and colleagues (2010) refer to as plot and theme analyses and/or holistic analyses. Plot and theme analyses utilize binary decision schemes to indicate whether or not children include specific story grammar components in their narratives and then quantify the number of target plotlines and themes they produce. Holistic analyses, on the other hand, rely on examiner judgment of narrative proficiency based on the overall quality and developmental level of the narrative. These typically rely on holistic rating scales, for which narrative proficiency is quantified on a continuum beginning with a non-narrative (e.g., simple labels or isolated actions) and ending with a complete and well-developed, or mature, narrative. Holistic scales are based on developmental data that describe the characteristics of the stories of typically developing, monolingual children at various ages (Glenn & Stein, 1980; Hughes et al., 1997; Peterson & McCabe, 1983; Stein & Glenn, 1979; Westby, 1984).

Language Productivity and Organization (Microstructural) Skills

Children's narratives constitute naturalistic language samples that allow measurement of their narrative organization skills as well as their lexical productivity. Measures of language productivity, or the amount of language produced by children, have been useful in establishing normative developmental perimeters for the expressive language development of both monolingual and bilingual children (Bedore & Peña, 2008; Miller & Iglesias, 2008). Language productivity is quantified in a variety of ways, including number of utterances or clauses (NU), total number of words (TNW), and number of different words (NDW). Sentence organization is most often measured by mean length of utterances (MLU) and the percentage of utterances that are grammatically

correct. Sentence complexity can be measured by subordination indices (SI), which measure the ratio of all clauses, including embedded or subordinate clauses, to main clauses.

While the focus of the current study is primarily on describing the narrative organization, or macrostructural skills, of SS-ELLs, analysis includes measurement of the following narrative microstructural elements: total number of words, number of mazes (e.g., false starts/hesitations), number of net words (total words minus mazes), and number of utterances.

Narrative Organization (Macrostructural) Skills

At the local level, children must attend to choice of words, sentence structure, and referential cohesion while, at the global or conceptual level, they must maintain an awareness of the overall meaning or gist of the story, its overarching theme, and its structure (Westby, 1992). Additionally, they must remain mindful of the listener's needs for information while engaged in what is essentially a social monologue (Owens, 2010). The assessment of children's narrative skills thus involves an evaluation of children's knowledge of content schema, or the relationships between story elements such as actors, settings, and events, and their knowledge of story grammar structure, through which children organize content schema in a coherent manner such that events are linked temporally and causally (Owens, 2010; Westby, 1992). Numerous models for story structure or story grammar have been developed; however, Stein and Glenn's (1979) is the most frequently used model for analyzing children's fictional narratives (Hughes et al., 1997).

Stein and Glenn (1979) developed an empirically based schema for story organization. They experimentally compared the content of children's recall,

comprehension and judgment of story information and found that, although some notable age and task differences were apparent, certain elements appeared to be stable across ages and tasks. They concluded that these stable elements constitute a sort of internal structure for the organization of stories. This hierarchical story structure takes as its unit of analysis the episode, which consists, minimally, of the introduction of a setting and characters plus an initiating event or internal state (the posing of a problem or a desire), goal-oriented action (an attempt to solve the problem or attain the desire), and a consequence (success or failure of the attempt). Story structures are represented hierarchically by the relationships (simultaneous, temporal or causal) between episodes. Variations of Stein and Glenn's story schema are commonly employed in narrative research (Hughes et al., 1997).

Story grammar analyses typically examine stories in one or both of two ways (Heilmann et al., 2010; Hughes et al., 1997): (a) Story transcripts are coded for instances of story grammar elements such as setting, initiating event or problem, resolution, etc. and frequencies of each element in a child's story are then described; (b) Stories are judged holistically for overall organization and coherence relative to a child's developmental level. The second approach relies on the examiner's judgment in assigning scores based on an ordinal rating scale designed to reflect a developmental hierarchy of critical story elements and episodic relations; thus higher ratings equate with better, more complete, and more mature stories (McFadden & Gillam, 1996).

Potentials and Limitations of Existing Methods of Narrative Assessment

Children's narratives constitute naturalistic language samples through which not only the quality of their oral narrative organization skills but also the amount of language they produce can be measured. Both types of measures have been useful in establishing

normative developmental perimeters for children's expressive language development (Bedore & Peña, 2008; Miller & Iglesias, 2008). While story grammar analyses have differentiated groups of monolingual children by developmental age and ability (Reilly, Losh, Bellugi, & Wulfeck, 2004), results vary widely across studies (Boudreau & Hedberg, 1999; Pearce, McCormack, & James, 2003; Reilly et al., 2004). This variation may result in part from differences in the storytelling tasks and elicitation procedures used by researchers (Fiestas & Peña, 2004; Pearce, 2003). It may also be attributed to the sensitivity of the scoring systems used to evaluate narrative skills. Heilmann and colleagues (2010) noted a ceiling effect when comparing the distribution of scores generated by three different narrative organization measures applied to a sample of children's narratives. The measures they applied were a plot and theme approach described in Reilly et al. (2004), a text-level measure that Manhardt & Rescorla (2002) adapted from Applebee (1978), and Pearce et al.'s (2003) ordinal adaptation of narrative levels described by Stein (1988). The findings, which they contrasted with results of their own protocol, the Narrative Scoring Scheme (one of the scoring systems used in the current study), were not surprising to the authors, who suspected that "existing narrative organization measures may focus too much on early developing narrative skills, such as the inclusion of key story grammar components" (p. 610). They suggest, "existing story grammar measures may be too easy and potentially insensitive for preschool and young school-age children" (p. 608). In addition to story grammar components, their measure thus incorporated higher-level, literate language skills related to mental states and character development, as well as referencing and cohesion. The inclusion of higher-level skills and the wider scale (0 to 35 points) of the measure resulted in a more normal distribution of scores for their school-age sample (ages 5 to 7), suggesting improved

sensitivity of the instrument to detect developmental differences among school-age children (Heilmann et al., 2010).

Lexical measures together with story grammar analyses can help to distinguish the narratives of typically developing children from those at risk for language and literacy-related problems. Compared with their typically developing peers, children with language and literacy difficulties are more likely to produce narratives that include less information and contain less grammatical complexity, lexical diversity, cohesive adequacy, and organizational coherence (Fazio & Naremore, 1996; Gutiérrez-Clellen, 1995; Hayward, Gillam, & Lien, 2007; Kaderavek & Sulzby, 2000; Paul & Smith, 1993; Roth, Speece, Cooper, & De La Paz, 1996). With respect to measures of story grammar, children with learning disabilities have been found to recall fewer events and are less likely to include characters' internal responses when reconstructing a story, especially as the stories they are asked to reconstruct become increasingly complex (Montague, Maddux, & Dereshiwsky, 1990; Ripich & Griffith, 1988).

STATEMENT OF THE PROBLEM

The current study aims to address two distinct but related problems pertaining to narrative assessment with ELLs. First, there is a substantial body of research (Celinska, 2004; Haynes, Haynes, & Strickland-Helms, 1989; Klecan-Aker & Kelty, 1990; Montague et al., 1990; Reilly et al., 2004; Ripich & Griffith, 1988; Roth & Spekman, 1986; Wolman, van den Broek, & Lorch, 1997) that describes and compares the narratives of typically developing, monolingual English-speaking children as well as monolingual children with language impairments and learning disabilities at various age levels. Considerably less research, however, has sought to describe the English narratives of Spanish-speaking English language learners (ELLs). Given that SS-ELLs

represent the largest language minority group in the country, a better understanding of their oral narrative skills in English provides a starting point for research on the oral narrative skills of the general population of ELLs living in the U.S. Second, narrative assessments produce varied results depending on both the foci and scales of available measures of narrative skills. Further, narrative assessments have been developed and tested almost exclusively on monolingual populations. There is a need to describe and compare the outcomes of such measures when used with bilingual children who are in the process of learning English and to identify those characteristics of narrative scoring systems that yield the most instructionally useful information. Knowledge of such characteristics will lead to the design of an appropriate narrative skill measure to use with ELLs as an informal way of assessing their literacy-related oral language skills. Such a measure may then inform classroom instruction and intervention for SS-ELLs who are learning to read in English (Griffin, Hemphill, Camp, & Wolf, 2004); furthermore, it may provide critical information to problem-solving teams considering special education referral (Bedore & Peña, 2008; McCabe, Bailey, & Melzi, 2008).

This exploratory study describes results of analyses of 83 oral English stories told by 42 second grade Spanish-speaking ELLs using three different oral narrative scoring systems: story grammar analysis (Stein & Glenn, 1979), Narrative Assessment Profile (Bliss & McCabe, 1998; McCabe & Bliss, 2003), and Narrative Scoring Scheme (Miller et al., 2006; Heilmann, Miller, & Nockerts, 2010). The choice of scoring systems was motivated by the desire to compare results according to systems that examine similar but also distinct aspects of children's narratives. Story grammar was specifically chosen because of its prevalence of use in the research on children's fictional narratives. The Narrative Assessment Profile and Narrative Scoring Schemes were selected because they have measurement properties that overlap with those of story grammar but additionally

capture microstructural features of oral narrative discourse and have been used with culturally and linguistically diverse populations. The study fills an important gap in the research by adding to the knowledge base on the English narrative skills of SS-ELLs and the scoring systems used to measure them. Since the majority of investigations of narrative skills have been conducted with monolingual English speakers, we have a robust literature base on the characteristics of narratives for this group. This study will facilitate comparison of the English narratives of ELLs to those of native English speakers. This is an important contribution given that most ELLs are served in ESL programs (August & Hakuta, 1997; Kindler, 2002) and thus are taught in English. Results will provide options for analyzing story skills which have been shown to be important, not only in terms of developing academic English, but also as foundational skills for reading comprehension. Furthermore, findings may suggest which approach(es) are most appropriate for analyzing stories told by SS-ELLs and what adaptations may be needed so that narrative assessment results in instructionally relevant and reliable information.

RESEARCH QUESTIONS

The purpose of this study is thus to contribute to the research literature in four ways: 1) by describing the organization and production qualities of the English fictional oral narratives of Spanish-speaking ELLs; 2) by comparing the results generated from three different narrative scoring systems when applied to analysis of these children's oral narratives; 3) by identifying the stable features of those narratives that are rated consistently across narrative scoring systems; and 4) by identifying criteria for developing a high quality scoring system for analyzing and evaluating the English oral

narratives of SS-ELLs and that produce data that are useful in identifying instructional goals and planning instruction for this population.

The narratives analyzed in this study come from an extant data set. The research questions are thus specific to 2nd grade SS-ELLs. The research questions guiding the investigation are the following:

1. What are the characteristics of second grade Spanish-speaking ELLs' stories, orally narrated in English, using the following methods of analyses?
    a. Story grammar analysis
    b. Narrative Assessment Profile
    c. Narrative Scoring Scheme
2. How does each scoring system characterize the sample in terms of expected narrative performance according to its own criteria?
3. What are the distinguishing features of narratives whose scores are consistent (e.g., high, average, and low) across measures?
4. What features must a narrative scoring system have in order to provide teachers with quality information that will help them design instruction and interventions for SS-ELLs?

SIGNIFICANCE OF THE STUDY

This study fills an important gap in the research by focusing on the English oral narratives of Spanish-speaking English language learners. Until we have a better understanding of the characteristics of the English oral narratives produced by SS-ELLs and how to best analyze them, the results of narrative assessment will be difficult to

interpret. A narrative macroanalysis scoring system appropriately designed for SS-ELLs may provide an informal, authentic, and holistic measure of an SS-ELL's oral narrative discourse skills that may be used to reduce the confusion created by language assessments that measure only the surface elements of language performance. This study will contribute to our understanding of the qualities of the English oral narratives produced by SS-ELLs in the second grade and it will identify features of narrative scoring systems that may be useful for instructional planning in classroom settings, thus facilitating appropriate intervention and improved educational outcomes for this population.

SUMMARY

Spanish-speaking English language learners are a fast growing yet under-served segment of the U.S. school population. They have historically experienced low achievement and disproportionate representation in special education under the category of learning disabilities. It has been suggested that these patterns do not reflect the actual ability of this population, but rather a state of professional confusion perpetuated by a shared knowledge base that does not adequately enable educators to distinguish academic difficulties associated with language acquisition from those stemming from learning disabilities. In the present study, the oral stories collected from a sample of second grade SS-ELLs are described in terms of narrative macrostructure, allowing for a set of characteristics and a range of qualities across those characteristics to emerge. Three different scoring systems are used to analyze each story and findings are compared across systems. Those features of each system that generate the most instructionally relevant information about the oral English stories of SS-ELLs in the second grade are identified

and issues related to the process of analyzing ELLs' fictional stories are documented and discussed.

The present investigation contributes to the knowledge base by describing a) the English oral narrative skills of Spanish-speaking ELLs in 2nd grade, b) the differences resulting from applying each of three distinct narrative scoring systems, c) features of SS-ELLs' narratives that are rated consistently when different scoring systems are applied, and d) the kinds of instructionally useful information that can be reliably gathered through narrative analysis using each scoring system. Issues that threaten the reliability of narrative analysis using the scoring systems under consideration are discussed and recommendations for adaptions to existing instruments and/or design considerations for the development of new instruments are provided. As an exploratory study, no specific hypotheses are being tested with respect to the research questions. However, based on the findings of others (Heilmann et al., 2010; Miller et al., 2006; Muñoz, Gillam, Peña, & Gulley-Faehnle, 2003; Roth et al., 1996; Roth, Speece, & Cooper, 2002; Speece, Roth, Cooper, & De La Paz, 1999) it is expected that the three different scoring schemes will produce varying and at times discrepant amounts and types of information related to the characteristics of the children's narratives with implications for the collection and interpretation of narrative data to inform instructional decision making.

CHAPTER TWO

Literature Review

Public education has been given the imperative to improve academic outcomes for the fast growing yet underserved population of elementary-aged Spanish-speaking English language learners. In a school accountability climate that demands the narrowing of a perpetual achievement gap between language minority students and their monolingual English-speaking peers, the present investigation is both timely and highly warranted. Given the historic disproportionate representation of Spanish-speaking ELLs in special education and the well documented confusion that prevails in special education referral processes for this population (Harry & Klingner, 2006; Klingner et al., 2005; Klingner & Harry, 2006; Wilkinson, Ortiz, Robertson, & Kushner, 2006; Ortiz et al., 2012), it is critical for research to address gaps in the knowledge base related to our field's ability to distinguish language and literacy performance problems arising from limited English proficiency versus those resulting from learning disabilities (Klingner, Artiles, & Barletta, 2006). It is only by eliminating such gaps in the knowledge base that we can hope to reverse the pervasive trend of inappropriate special education referrals of culturally and linguistically diverse students in high incidence special education programs (Artiles & Klingner, 2006; Klingner et al., 2005). The current study aims to contribute to the closing of such gaps by describing the characteristics of second grade Spanish-speaking ELLs' oral narrative skills and by identifying scoring systems that are appropriate to the task of evaluating such skills.

This chapter reviews literature related to (a) the population of Spanish-speaking ELLs in the United States, specifically their prevalence, (b) their educational outcomes,

and (c) factors influencing their disproportionate representation in special education, specifically the process of and challenges associated with distinguishing reading difficulties arising from limited English proficiency versus those arising from reading disabilities, and the limitations of available assessments in helping to make those distinctions; and (d) considerations in conducting narrative assessments with SS-ELLs. Next, literature related to each of the methods of narrative assessment used in the current study is reviewed: Story Grammar Analysis; Narrative Assessment Profile; and Narrative Scoring Scheme. Finally, the results of a systematic review of the narrative skills of SS-ELLs are described.

THE GROWING POPULATION OF SPANISH-SPEAKING ENGLISH LANGUAGE LEARNERS

Census data indicate that, as of 2009, 21% of children ages 5-17 spoke a language other than English at home, and 5% spoke English with difficulty (Aud et al, 2011). These children are collectively known as language minority children or English language learners (ELLs) and the 5% who speak English with difficulty are typically identified as Limited English Proficient (LEP). Spanish-speaking English language learners (SS-ELLs) are the largest segment of language minority school-age children in the U.S. Numbering approximately 3.5 million, they constitute more than 80% of all ELL students in 14 states and the predominant subgroup of ELLs in 44 states and the District of Columbia (NCELA Fact Sheet, 2011). Nationwide, Spanish-speaking ELLs represent 73% of school-age children who speak English with difficulty (Aud et al., 2011). The Hispanic population is growing at nearly four times the rate of the total U.S. population and is projected to more than double in size from 2010 to 2050; currently at approximately 16%, it is expected to constitute 25% of the U.S. population by 2050 (U.S. Census Bureau, 2006). According to the 2010 census, the total population growth

between 2000 and 2010 by region averaged three to four percent for the Northeast and Midwest and about fourteen percent each for the South and West regions. However, during this same period, the Hispanic population increased between 33% and 57.3% across all regions. More than half of the total population growth in the US between 2000 and 2010 was due to the growth in the Hispanic population (Ennis, Riós-Vargas, & Albert, 2012). This population is concentrated overwhelmingly in California and Texas, in which reside 27.8% and 18.7% of the Hispanic population respectively. The Hispanic population is a relatively young population; in 2010 it represented an estimated 16.1% of the resident population of the U.S., but 26.1% of the population under 5 years old and 21.8% of the population ages 5-17 (Table 21, Snyder & Dillow, 2011). In 2008 in California, New Mexico, and Texas, the majority of children enrolled in public schools were Hispanic (Table 43, Snyder & Dillow, 2011).

SCHOOL RELATED OUTCOMES

Pervasive Achievement Gap

Given the rapid growth of the Hispanic population at or approaching school age, it is especially troubling that the achievement gap between White/non-Hispanic students and Hispanic students has been and continues to be so pervasive. In 2009, the total status dropout rate for all ethnicities was 8.1%. However, the dropout rate was 5.2% for Whites, 9.3% for Blacks, and 17.6% for Hispanics (Table 115, Snyder & Dillow, 2011). While the overall dropout rate has improved, dropping below 10% for the first time in 2003, the disparity in dropout rates between Hispanics and all other ethnicities has not improved. The dropout rate for Hispanics has remained two to three times that of the overall dropout rate since 1972 and three to four times greater than the dropout rate for Whites (Table 115, Snyder & Dillow, 2011). Considering the trends evident in

Department of Education data, it is not surprising that the dropout rates for Hispanics are so high; our schools clearly fail to close a gap that begins well before children enter school despite ample opportunity to do so during children's early, most formative years.

Consider these data published in the *Digest of Education Statistics 2010* (Snyder & Dillow, 2011). These data were collected as part of the Early Childhood Longitudinal Program, which used the Bayley Short Form Research Edition (BSF-R). At both nine months and two years of age the assessment was administered in the child's native language by a bilingual interviewer or with the assistance of an interpreter. At nine months of age, there appear to be no significant differences between ethnic or racial groups on the demonstration of specific cognitive skills or specific motor skills. Averages are fairly similar across all groups and all categories of skills. At two years of age, school-related cognitive skills such as receptive and expressive vocabularies, listening comprehension, matching/discrimination, and early counting are measured; here differences become pronounced. In each of these categories Hispanic children perform well below their White, non-Hispanic peers and well below the average for all children. By contrast, there are no appreciable differences between groups of two-year old children on specific motor skills, suggesting disadvantages are strictly education related. Comparing the levels of specific cognitive skills between socioeconomic groups results in similar gaps between the lowest 20 percent, the middle 60 percent, and the highest 20 percent. Hispanic children's performance, as with Black children's performance, is closest to the averages for the lowest 20 percent (Table 119, Snyder & Dillow, 2011). The differences at age two would suggest that, insofar as these measures of specific cognitive skills reliably predict academic success, children who are Hispanic, Black, or American Indian/Alaska Native and/or are from the lowest socioeconomic groups have school-related disadvantages well before they ever enter school. The situation does not

improve after age two. At age four, children's reading, language, mathematics, color knowledge, and fine motor skills are measured. At age five, reading, mathematics, and fine motor skills are again measured. At kindergarten, first, third, fifth, and eighth grade, reading, math, and science skills are measured. At all points and for all cognitive measures (but not motor skills), there is a persistent gap between levels of performance demonstrated by children who are White versus children from all other racial and ethnic groups, except Asian. This gap is also evident between socioeconomic groups (lowest, middle, and highest) and is present each reported year for the past four decades (Snyder & Dillow, 2011).

Given that children were more developmentally alike than different prior to their first birthday, one has to wonder, what is the source of disadvantage and why does it appear to predominantly and pervasively affect children who are Hispanic, Black, American Indian/Alaska Native and/or of the lowest socioeconomic status; and, more importantly perhaps, what can be done about it? These questions have dominated policy debates surrounding school reform for decades. Discussions about the achievement gap between White and minority students and the disproportionate representation of minority students in special education often revolve around the contributions of poverty (MacMillan & Reschly, 1998; Skiba, Poloni-Staudinger, Simmons, Feggins-Azziz, & Chung, 2005) and of multiple other factors that are beyond the influence of schooling to affect (Evans, 2005). However, many have maintained that the roots of the problem are much deeper, insidious, structural and hegemonic; they spring in large part from the unquestioned and unchallenged White, middle class cultural norms that govern the project of public education and the ways in which the institution of public schooling has systemically dealt with difference as deviance, deficit, and/or disability (Artiles, 2009; Artiles, 2011; Artiles et al., 2010; Artiles & Trent, 1994; Blanchett, Mumford, &

Beachum, 2005; Coutinho & Oswald, 2000; Cummins, 2001; De Valenzuela, Copeland, Huaqing Qi, & Park, 2006; Delpit, 2006; Deno, 1970; Donovan & Cross, 2002; Dunn, 1968; Eisner, 2003; Garcia & Guerra, 2004; Harry & Klingner, 2006; Klingner et al., 2005; Leone et al., 2003; Losen & Welner, 2001; Ogbu, 1992; Ogbu & Simmons, 1998; Patton, 1998; Rumberger & Larson, 1998; Skiba, Michael, Nardo, & Peterson, 2000; Skiba et al., 2008; Trueba, 1988; Weinstein, Gregory, & Strambler, 2004). An ecological inspection of this problem implicates cultural biases operant at multiple levels of education (Klingner et al., 2005), including teaching (Garcia, Arias, Murri, & Serna, 2010; Garcia & Guerra, 2004; Marx, 2004; Villegas & Lucas, 2002), learning and cognition (Cole, 1996; Rogoff, 2003; Rogoff & Chavajay, 1995), assessment and evaluation of students (Abedi, Hofstetter, & Lord, 2004; Cummins, 2001; Figueroa, 2002; Garcia, 2002; Ortiz & Yates, 2002), and expectations related to student behavior (Krezmien, Leone, & Achilles, 2006; Skiba et al., 2000), performance (McKown & Weinstein, 2008; Weinstein et al., 2004), and the relationships between communities and schools (Cummins, 2001; Ogbu, 1992). When considered from a sociocultural/ecological perspective, the issues contributing to educational inequity are highly complex and well beyond the scope of this literature review to address thoroughly. Nevertheless, they point to a need to deepen our understanding of the ways in which culture and language influence learning and school performance, something the current study seeks to address by increasing our understanding of Spanish-speaking children's English oral narrative performance.

The trends contributing to the persistent achievement gap described in the preceding paragraphs suggest that, just by virtue of being a Hispanic in the U.S., SS-ELLs are at risk of school failure or, at the very least, underachievement. However, not all Hispanic children are English language learners or Limited English Proficient. For

Hispanic children who are in the process of acquiring English proficiency and who are not served by quality ESL or bilingual education programs, the risk increases exponentially. In schools, ELLs who are labeled as Limited English Proficient (LEP) are concentrated in the early elementary grades (K-3) with enrollments steadily decreasing in succeeding grades (Kindler, 2002). They are a very heterogeneous group in that they enter school with a wide range of demographic characteristics such as immigration and socioeconomic status, as well as school-related experiences. English language learners who are immigrants differ in age of arrival in the U.S., educational background, language and literacy proficiencies in English and in their native languages, and subject matter knowledge. Students who were born and have been raised in the U.S. but who speak a language other than English at home differ in their levels of native language and literacy proficiency as well as proficiency in English. Historically, the academic achievement of ELLs has lagged well behind that of their native English-speaking peers; they are more likely to repeat a grade, to be placed in lower ability groups, and to drop out of school (August & Hakuta, 1997; Klingner et al., 2008). English language learners fare especially poorly in English reading. According to the Digest of Education Statistics, nationwide in 2009, only 29% of ELLs in the 4th grade met basic achievement levels in reading while only 6% met proficient levels. In 8th grade, 25% of the ELLs who were assessed met basic reading achievement levels and 3% were proficient (Table 132, Snyder & Dillow, 2011). By comparison, 69% of 4th grade students who were not ELLs achieved at the basic level and 34% at the proficient level while 76% of non-ELL 8th graders at or above basic reading levels and 32% were at or above proficient. Because these data are about English reading without consideration for important factors such as language proficiency or language of instruction, the results should be interpreted with caution.

Disproportionate Representation in Special Education

There has been concern about special education identification rates for this population as well (Artiles, Rueda, Salazar, & Higareda, 2005; Artiles & Trent; 1994; Donovan & Cross, 2002; Zehler, Fleischman, Hopstock, Pendzick, & Stephenson, 2003). The proportion of students with disabilities whose home language is not English has increased substantially along with the LEP population. However, special education identification rates vary considerably from district to district, reflecting patterns of both over- and underrepresentation of LEP students in special education. Zehler and colleagues (2003) conducted a nationally representative study of LEP students enrolled during the 2000-2001 school year, including those served by special education. They found differences in special education identification rates related to the number of LEP students enrolled in a district. In the general population, 13.5% of all students received special education but only 9.2% of the LEP population were identified as needing special education services. For those districts serving higher numbers (100 or more) of LEP students, special education identification rates were lower than they were for the general population (9.1% as compared with 13.5%) suggesting LEP students may be under-identified as having special education needs. However, districts serving smaller LEP populations, defined by Zehler and colleagues as fewer than 100 LEP students, had higher identification rates. In such districts an average of 15.8% of LEP students received special education services. They suggest that further research is needed to explore the reasons for such variation and noted "one factor is the difficulty encountered by staff in assessing LEP students and in distinguishing second language acquisition versus a disability" (p. 28). The authors additionally report the types of services received by LEP students identified as needing special education (Sp-ED-LEP). Contrary to a trend toward inclusion in special education, Sp-ED-LEP students are educated in separate

settings more frequently than are special education students in general. Their services appear to be mostly uncoordinated mixtures of Special Education and LEP services, resulting in less extensive LEP services and less native language instruction than are received by their LEP peers without special education needs. Clearly, any discussion of performance and placement patterns for this population needs to consider the language of instruction and the overall quality of instructional programming made available to them.

While Zehler and colleagues confirmed differences in placement patterns associated with district size, Artiles and colleagues (2005) found similar patterns associated with student level variables such as grade level and language proficiency status. They examined special education placement patterns of ELLs in eleven urban districts in southern California. The population of these districts in aggregate was majority Latino/a and 42% of the student population was classified as ELL. Over 90% of the ELL population was Spanish-speaking. The districts classified two types of ELLs: those with limited proficiency in English and those with limited proficiency in both English and their first language. The proportion of ELLs receiving special education services was similar to that of the general population, at around 7.6%. However, there were pronounced differences associated with grade level and immigration status. Artiles et al. compared patterns of special education placement for subgroups of the ELL population and subgroups of the general, non-ELL population. Looking at grade level groups, ELLs were overrepresented at the secondary level, whereas English proficient students were underrepresented. At the elementary level, ELLs appeared to be underrepresented. Another trend was evident when looking at ELLs across grades. Whereas the population of ELLs decreased across the elementary grades, the proportion of ELLs in special education increased across the elementary grades, shifting the pattern of underrepresentation in Grades K-5 to overrepresentation in Grade 6. Secondary aged

ELLs with limited proficiency in both L1 and L2 were significantly overrepresented in the three disability categories of mental retardation (MR), language and speech impairments (LAS), and learning disabilities (LD) while ELLs with limited proficiency in L2 were overrepresented in the LD category; all other groups were underrepresented. Likewise, elementary aged ELLs with limited L1 and L2 proficiency were overrepresented in the LAS and LD categories, whereas, again, most other groups were underrepresented. Given the great discrepancies between special education placement patterns of subgroups of ELLs in these districts, Artiles and colleagues emphasize the importance of including population subgroups in analyses examining disproportionality. They also note the need to better understand the specific characteristics of students who are considered to be limited in both their L1 and L2 and to examine the factors contributing to overrepresentation of this population in certain special education disability categories. Their findings support the need for the present investigation, which seeks to better understand the characteristics of the English oral narrative skills of SS-ELLs and to identify appropriate ways to assess those skills.

Others have similarly suggested special education placement patterns for LEP students reflect confusion among district personnel as to whether LEP students' academic difficulties especially in the area of literacy result from the language acquisition process or from disability. Harry and Klingner (2006) describe the challenges one large, culturally and linguistically diverse urban district faced in addressing the problem of inappropriate referrals of large numbers of CLD students for special education. They found that, although the district made a good attempt to protect against such inappropriate referrals on paper, it was evident that such intentions did not often translate to practice. Many factors, including the widely variant knowledge and skills possessed by child study team members, contributed to inconsistent implementation of pre-referral processes

designed to lead to more equitable and appropriate decision making. Specifically, with respect to bilingual children, district referral policies and guidelines for ELLs were excellent according to Harry and Klingner. Nevertheless much variability prevailed in the actual pre-referral and referral processes observed by the research team. Some of the barriers to implementing these otherwise sound written guidelines included inadequate assessment procedures reflected in the absence of bilingual assessors at meetings and placement conferences, confusion regarding the roles, responsibilities and expertise of various key staff members, and difficulties differentiating between normal second language acquisition and learning disabilities. These barriers resulted in variable placement patterns within the district and the existence of both high- and low-referring schools within the district. Upon closely examining child study team, multidisciplinary teams, and referral processes for ELLs, the Klingner and Harry (2006) reported much confusion about when to refer an ELL for an evaluation and when to conduct assessments in English. In the absence of such clarity, teams tended to rely heavily on the opinions of parents and teachers about the level of a child's English proficiency and his or her readiness to be tested in English. Furthermore, they noted language issues were poorly considered when interpreting students' achievement difficulties resulting in limited English proficiency being interpreted as low IQ or learning disabilities. An equally problematic assumption observed by Klingner and Harry was the infallibility of psychological evaluations in accurately diagnosing disabilities. Likewise, the psychologist was given the most authoritative role in decision-making processes, and very little effort was made to design meaningful pre-referral intervention strategies prior to conducting a formal evaluation. Other problems observed include confusion regarding the role of the bilingual assessor, decisions largely made prior to placement conferences, and poor or inadequate efforts to include parents meaningfully in decision-making

processes. The current study aims to reduce confusion in decision-making processes by contributing to our knowledge of patterns of English oral narrative performance.

Wilkinson, Ortiz, Robertson, and Kushner (2006) sought to develop profiles of Spanish-speaking ELLs who had received literacy instruction in their native language and whose identification as having learning disabilities was based on documentation of performance in that language. As a component of this endeavor, an expert panel reviewed the documentation that had been considered by the schools' multidisciplinary teams (MDT) in determining each child's eligibility for special education services for reading under the category of learning disabilities. This effort was initially undertaken to validate the appropriateness of the MDTs' decisions before developing profiles. However, based on the wide variation encountered in the sample, the researchers decided to focus on 21 students who were classified as having an LD in reading with no secondary impairments at the time of initial entry into special education and who continued to be served solely under the category of LD at the time of the actual study. Each member of the expert panel reviewed each of the student records and determined whether, in her opinion, the student under consideration would qualify for special education for a reading-related disability. Furthermore, each panelist specified which factors in the students' files led to such a conclusion; what types of information or processes were expected but not present in the folders; and whether or not the data presented were sufficient to support a determination of eligibility in light of the exclusionary clause. After reviewing student data individually, decisions were compared. Decisions were unanimous for 13 of the 21 students. For the remaining 8, the panel discussed available data to reach a consensus. Descriptions of factors important to the decisions were generated and these, along with the eligibility decisions themselves, were compared between the district personnel's decisions and those of the expert panel. The

district identified all 21 of the students in the sample as meeting criteria for having a learning disability using the state's discrepancy formula, which requires a documented discrepancy between a student's IQ and his or her achievement. The state also allows the MDT to base an eligibility decision on other evidence if it is determined that a discrepancy cannot be established because of a lack of appropriate instruments; however, this alternative method was not employed in any of these cases. Although the district based its eligibility decisions exclusively on discrepancy criteria, the specific procedures it employed were difficult to summarize due to wide variation. Upon comparison, the expert panel agreed with the LD eligibility determination for only 11 of the 21 students. For the other 10 students, the panel determined that there were sufficient other factors to which learning difficulties could be attributed and that additional data would be necessary to make an eligibility determination. Furthermore, the expert panel disagreed with the district's diagnosis of reading-related learning disabilities for 6 of the 11 students for whom the LD classification was deemed appropriate. Therefore, slightly less than 25% of the district's eligibility and classification decisions were validated by members of the expert panel, who then described those factors constituting evidence that learning difficulties were best explained by a reading-related LD. Those factors include consistent school histories, substantially low achievement scores in Spanish reading, and the presence of reading difficulties over time despite specialized interventions in general education. For these children deemed to have a reading-related LD, language of assessment was consistent with language of instruction and their files contained multiple indicators of reading difficulties that could not be attributed to other factors, such as an attention deficit disorder or a head injury. Even though the expert panel disagreed with the district's decision to classify 15 of the 21 students as having a reading-related LD, there were legitimate documented concerns regarding all of the children, including

concerns reported by parents. Nevertheless, the panelists noted numerous procedural problems that limited the ability of the MDT to make appropriate decisions regarding classification. Although based on a very small sample size, these findings corroborate those of larger studies (Artiles et al., 2005; Harry & Klingner, 2006; Klingner & Harry, 2006; Zehler et al., 2003) and provide valuable insight into factors contributing to inappropriate special education placement of ELLs, especially in the category of reading-related learning disabilities. They additionally point to a need for continued research to develop our understanding of the characteristics of ELLs with reading-related LD and how to identify LD in this population, ruling out limited English proficiency as the cause of reading difficulties.

FACTORS CONTRIBUTING TO INAPPROPRIATE SPECIAL EDUCATION OUTCOMES

Distinguishing Reading Difficulties Arising from Limited English Proficiency from Those Arising from Reading Disabilities in Spanish-Speaking English Language Learners

> Identification of English language learners with learning disabilities is hampered by a lack of theory and empirical norms that describe the normal course of language and literacy development for English language learners and the individual, school, and social factors that relate to that development. The context provided by profound differences in the nature of prior schooling cannot be ignored. One of the reasons for limitations in existing knowledge is that some necessary studies require the availability of comparable assessments or language-general identical assessments, neither of which has been available until recently. (Wagner, Francis, & Morris, 2005, p. 13)

The challenges associated with distinguishing learning disabilities from language acquisition processes in ELLs are well documented (August & Hakuta, 1997; August & Shanahan, 2006; Artiles & Klingner, 2006; Klingner et al., 2008; McCardle, Mele-McCarthy, & Leos, 2005; Ortiz, 1997; Wagner et al., 2005). Although the current study is not investigating reading or the process of identifying reading disabilities in SS-ELLs, it contributes to our knowledge of empirical norms with respect to literacy-related oral

language development. A brief review of literature addressing these challenges is thus included.

Wagner and colleagues (2005) discuss the advantages of adopting a social systems/dimensional perspective of LD versus the prevailing medical/categorical model. Learning disabilities, they maintain, are not well characterized by a medical model/categorical perspective, which relies on the presence or absence of criteria; rather, learning disabilities are characterized by a continuous and multivariate distribution of performance. The only way to classify presence or absence of a learning disability from a categorical perspective is to establish cut-points in the distribution. It is notoriously difficult to specify and to validate such cut-points because, unlike medically diagnosable low-incidence disabilities, learning disorders are socially and linguistically moderated; criteria differ from state to state and are dynamic across the lifespan, leading to instability of classification. Faulty, precarious, or unstable classification of LD carries implications for treatment and outcomes for individuals. Wagner and colleagues thus emphasize "in considering treatment effectiveness, we must get beyond group mean differences to consider for whom treatments are effective" (p. 9). The identification issues presented and discussed are complex even for the monolingual, English-speaking population of students. They are much more complex, however, for the population of ELLs. It is necessary, the authors maintain, but not sufficient, to develop comparable assessments for ELLs in both their native language and in English in order to gain a more complete understanding of a student's knowledge, skills, and instructional needs. Beyond comparable assessments, sociocultural and linguistic variables complicate the process of identifying learning disabilities in this population. Research, such as that which the current study proposes, is needed to understand normative language and literacy

development for ELLs from various linguistic backgrounds and also to understand how learning disabilities present within this population.

Klingner, Artiles, and Barletta (2006) reviewed empirical research on ELLs with reading difficulties as well as ELLs with learning disabilities. Their purpose was to identify research indicators that may help to better differentiate ELLs who struggle with reading due to limited English proficiency from those who struggle because they have a reading-related learning disability. After selecting published research that addressed ELLs who struggle with reading, they reported findings thematically. Specifically, findings were reported under the following categories: a) population subtypes, b) the role of context in understanding ELLs who struggle to read, c) issues pertaining to prereferral and referral, d) assessment practices with ELLs who may have an LD, e) predictors of reading achievement, f) instructional interventions, and g) ways in which literacy acquisition processes in a first and a second language can inform LD identification. Their overriding conclusion was that much more research is needed, specifically research in which ELL participants are described in much greater detail than is commonly the case. Existing evidence indicates that some subpopulations of ELLs are more vulnerable to special education placement than are others, but not enough research describes these various subpopulations, particularly with respect to levels of language proficiency in their native languages and in English. Similarly, they emphasized the need to better understand the roles of language and culture in assessment practices and to devote research to developing detailed profiles of ELLs who struggle with literacy. They further stressed the importance of assessing ELLs' strengths in alternative ways, such as the narrative assessment methods used in the current study, and considering numerous ecological, cultural and affective factors.

One ecological factor that must be considered when attempting to determine whether reading difficulties occur due to language acquisition or to learning disabilities is the quality of the educational environment itself. Ortiz and colleagues (Artiles & Ortiz, 2002; Garcia & Ortiz, 2006; Ortiz, 1997; Ortiz, Wilkinson, Robertson-Courtney, & Kushner, 2006; Ortiz & Yates, 2001) elucidate contextual and systemic factors in educational environments that may serve to hinder or facilitate the effective education of ELLs and the appropriate identification of ELLs with disabilities. Attention to these factors is required if we are to reduce or prevent inappropriate referrals of ELLs to special education. Effective school environments for ELLs prioritize: a) prevention and early intervention of learning problems, b) referral processes that take into consideration relevant and multiple data (including data from authentic, informal assessments such as the oral narrative assessments used in this study) and that minimize bias, c) assessment processes conducted by qualified bilingual evaluators, d) multidisciplinary teams composed of professionals with expertise in the education of ELLs, e) IEPs that are culturally and linguistically relevant, f) instructional programming in the least restrictive environment that addresses both disability-related and language needs, and g) annual reviews that evaluate progress and update language proficiency and dominance data (Ortiz & Yates, 2001). To the extent that ELLs are educated in school environments that don't prioritize these recommended practices, we cannot rule out the possibility that learning difficulties are attributable to the lack of appropriate, culturally and linguistically relevant instruction and assessment. Prevention is therefore key and begins with the school and classroom contexts that promote an additive approach to cultural and linguistic diversity (Cummins, 2001), collaboration between schools and the communities they serve, academically rich programs, and highly skilled teachers. When concerns about academic performance arise despite sound preventive contexts, culturally and

linguistically responsive assessment practices are paramount to making appropriate eligibility decisions.

Culturally and Linguistically Responsive Assessment Practices for ELLs

In order to develop culturally and linguistically responsive assessment practices to ensure the accurate identification of LDs in ELLs, several key research gaps need to be addressed. McCardle and colleagues (2005) discussed the themes emerging from the October 2003 National Symposium on Learning Disabilities in English Language Learners, at which research priorities and needs were discussed. The five major themes generated at the symposium were (1) identification and assessment of learning disabilities/reading disabilities, (2) understanding the language/literacy developmental trajectories of ELLs, (3) understanding individual and contextual factors affecting outcomes, (4) the intersection of each of these areas with neurobiology, and (5) developing and empirically validating effective interventions for LD in ELLs. With respect to identification and assessment of learning and reading disabilities in ELLs, there is a need to better understand how specific LDs will manifest in different languages. The development and validation of a theory-driven classification system of LDs in ELLs was thus emphasized.

With respect to the latter, Ortiz and Yates (2002) suggest that comprehensive language evaluations of ELLs ought not to rely exclusively on norm-referenced instruments, which provide incomplete profiles of language skills, but must incorporate language samples collected under more naturalistic conditions and which provide information about a range of language skills. They recommend using storytelling tasks to provide insight into children's narrative skills, which include the ability to organize and sequence information, draw conclusions, and evaluate actions.

CONSIDERATIONS IN CONDUCTING NARRATIVE ASSESSMENTS WITH ELLS

Care must be taken when interpreting the results of narrative assessments with ELLs as there is scant research available to guide the interpretation of narrative performance. A synthesis of research on the narratives of school-age Spanish-speaking English language learners revealed that, while their narratives do reflect typical developmental patterns such as amount and complexity of language, they also are characterized by performance patterns more typical of monolingual children with language and/or learning disabilities (McFarland, 2011). Given storytelling tasks, children with LD tend to perform poorly on: topic maintenance (McCord & Haynes, 1988); cohesion (Montague et al., 1990; Ripich & Griffith, 1988; Roth, Spekman, & Fye, 1995); amount of information provided (McCord & Haynes, 1988; Montague et al., 1990; Roth & Spekman, 1986); proportion of complete episodes (Montague et al., 1990; Roth & Spekman, 1986); inclusion of consequences, settings and internal responses (Montague et al., 1990; Ripich & Griffith, 1988); organization (Montague et al., 1990); inclusion of responses, attempts, and plans (Ripich & Griffith, 1988; Roth & Spekman, 1986); causal and concurrent relations (Roth & Spekman, 1986); and inclusion and accuracy of events in story recall tasks (Ripich & Griffith, 1988).

These patterns may underscore the influence of emergent bilingualism on the cross-linguistic narrative skill sets of the Spanish-speaking ELLs (Bialystok, 2007). Such variable performance patterns reflect the dynamic relationship between the process of English language and literacy acquisition and the individual differences characteristic of English language learners in the U.S., where differences in native language proficiency, previous learning, quality of instruction, and cognitive abilities contribute to increasingly variable performance patterns and achievement gaps over time (August & Shanahan, 2006). As a result, when compared to the narratives of monolingual English

speakers of similar age, the narratives of typically developing ELLs may resemble those of same-age monolingual speakers with disabilities or of younger, typically developing monolingual speakers. As a group, ELLs are likely to demonstrate wide variability in narrative micro- and macrostructure in both English and Spanish, but especially in English, over time. This may result in much confusion when evaluating ELLs who struggle with reading and lead to inappropriate or untimely interventions and instructional arrangements. Specifically, it may lead to the over- or under-identification of ELLs with reading-related learning disabilities (Artiles & Ortiz, 2002; Klingner et al., 2008).

The Need to Investigate the Narratives of ELLs across the School Years and in Relation to Reading

Research investigating the narratives of monolingual children with and without disabilities has paid special attention to the traits of such narratives at different ages, including preschool, lower elementary, upper elementary, middle school, and even high school aged youth. Both cross-sectional and longitudinal research spanning the school years have allowed some normative patterns of narrative performance and its relation to reading performance within this population to emerge. By contrast, most studies of the narratives of ELLs focus narrowly on younger children in kindergarten through 3rd grade (McFarland, 2011). Likewise, while studies on the narratives of monolingual children with learning disabilities have been published, those focusing on the narratives of ELLs with LD are virtually nonexistent; rather the focus is on ELLs considered to be typically developing or those with speech and language impairments.

Research on monolingual children with and without disabilities indicates that the qualities of children's narratives change with age, that these age-related changes differ between groups with and without LD, and that the relationships between narrative

performance and reading comprehension change over time and also differ in some ways between children with and without language- or reading-related disabilities (Dickinson & McCabe, 2001). For example, Snyder and Downey (1991) compared the narrative skills of eight- to fourteen-year-old children with and without reading disabilities and examined the relationship of their narrative skills to their reading comprehension. The children's performance on two story retelling tasks differed significantly between the two groups. Additionally, the older children in each group performed significantly better than the younger children. Stepwise multiple regression analyses revealed that the narrative scores of the children without disabilities predicted a significant amount of variance in their reading comprehension scores and that the amount of variance accounted for increased with age. By contrast, different variables predicted reading scores for the younger and older group with reading disabilities (RD), who differed from the typically developing children in both single-word decoding and silent reading comprehension. When standard scores were used, the single-word decoding and silent reading comprehension scores of the typically developing younger and older children did not differ. However, an age effect was evident for the RD group, whose single-word decoding scores remained positionally the same on the curve over time while their reading comprehension scores improved. In other words, the decoding skills of children with RD did not significantly improve, nor did they account for improvements in reading comprehension. Rather, sentence completion, naming speed, and naming accuracy predicted reading scores for the younger children with RD, while narrative discourse inference was the single greatest predictor of reading comprehension for the older children with RD. The authors interpreted these findings to suggest that, while children with RD retain decoding-skill deficits as they mature, they learn to compensate by using discourse-processing skills to aid their reading comprehension.

Adlof, Catts, and Lee (2010) conducted a longitudinal study of language and reading development with 433 children who were followed from kindergarten to eighth grade. Various oral language skills, including narrative expression and comprehension, were measured in kindergarten while reading comprehension was assessed in second and eighth grade. Medium to moderately high correlations were found with each of the measures and reading at both times. The strongest correlations between kindergarten and second grade measures were letter identification and sentence imitation while the strongest correlations with eighth grade reading were sentence imitation and grammatical completion. Logistic regression analyses revealed that no single measure was able to optimally predict reading comprehension at the two grade levels. Best-fit models were generated to predict dichotomous reading comprehension status (good and poor readers) at each grade level. The single best predictor at second grade was letter identification and the best model included sentence imitation, letter identification, mother's education level, rapid naming, phoneme deletion, narrative comprehension, nonverbal IQ, and picture vocabulary. The most important predictors of eighth grade reading comprehension status were phoneme deletion, grammatical completion, nonverbal intelligence, sentence imitation, mother's education level, narrative expression, narrative comprehension, and oral vocabulary. The authors note that current screening batteries, which focus on phonological awareness and alphabet knowledge, may fail to detect children who are at risk for reading comprehension difficulties in the later grades. Including a broader array of oral language and cognitive skills measures in early screening practices may aid in the early identification of poor comprehenders.

Cross-Linguistic Relationships between Narrative Performance and Reading Performance of ELLs

Miller et al. (2006) investigated the relationship between oral language and reading in a group of 1,531 Spanish-English bilingual children in kindergarten through third grade. Oral narrative samples were collected and analyzed for all children in both languages. Miller and colleagues examined the relationship of each of the narrative measures to reading comprehension and word reading efficiency in each language. Using regression analyses, they found that oral language measures, which included MLU (mean length of utterance), NDW (number of different words), WPM (words per minute), and NSS (narrative structure score), predicted reading scores both within and across languages beyond the variance accounted for by grade level. Of all the oral language measures, only NSS was initially stronger in Spanish than in English and remained stronger up until the 3rd grade. Oral language skills contributed to more variance in the passage comprehension than they did to the word reading efficiency scores and English oral language measures accounted for more variance beyond grade level in English reading than did Spanish oral language measures in Spanish reading. Likewise, English oral language measures contributed more unique variance (6%) to Spanish reading comprehension than did Spanish (2%) to English reading comprehension. The authors attributed this finding to more classroom time spent in English instruction as well as the increased exposure to and use of English within the children's communities, which were in Texas. The authors concluded that better oral language skills appeared to have facilitated reading in either language. Furthermore, while relationships between oral language and reading were strongest within languages, there was ample evidence of cross-linguistic influence. This underscores the need "to examine the child's performance in both languages to get the most complete picture of the student's strengths

and weaknesses and the full linguistic resources that the child is able to bring to bear in performing academic work" (p. 40). They stress that implications for understanding reading disabilities in bilingual children include paying particularly close attention to native language deficits in the preschool years, as these may be considered risk factors for both language-based learning disabilities and reading disabilities considering the role that oral language plays in the latter.

Had Miller and colleagues included older ELLs or ELLs with disabilities in their sample, they may have found patterns of oral language skills, including narrative skills, predicting reading comprehension differentially by age level and by disability status. In the only study of its kind known to this author, Goldstein, Harris, and Klein (1993) examined the relationship of oral storytelling and reading comprehension in a group of older Latino students with learning disabilities. Similar to Snyder and Downey (1991), they found a significant, moderate positive correlation between the story structure analyses and the reading comprehension scores of their junior high school subjects, all of whom were native Spanish speakers with previously identified LEP status. Their study was conducted solely in English and so no conclusions can be drawn with respect to cross-linguistic relationships between story structure and reading comprehension. The authors also cautioned that their story structure measure was adapted from a standardized instrument of oral language production and was not validated.

Gaps in the Research on the Narrative Skills of ELLs

The corpus of research published on the narratives of ELLs leaves many gaps (McFarland, 2011); these include: 1) insufficient descriptions of the samples; 2) lack of studies investigating the narrative skills of ELLs with learning disabilities and the relationship of those skills to reading comprehension across the school years; and 3)

incomparable findings between studies due to important methodological differences in tasks, elicitation procedures, type and amount of contextual support for storytelling, and the researchers' criteria for coding, analyzing, and evaluating narratives. The current study seeks to remedy the last stated gap by comparing findings generated by a common elicitation task and identifying appropriate methods for coding, analyzing, and evaluating narratives.

Insufficient Descriptions of the Samples

ELLs are an inherently diverse population given a common label by virtue of a single shared characteristic: their status as bilinguals with less than native like proficiency in English. When conducting research with and on ELLs, it is therefore crucial to describe the sample sufficiently to account for substantial within-group variability (Artiles & Klingner, 2006). However, sample characteristics are rarely reported in sufficient detail (McFarland, 2011). Immigration generational status, national origin, SES, parent education, and previous schooling experiences constitute major sources of variation within the U.S. population of ELLs with important ramifications for public education. Recent immigrants are both more educated and less educated than native-born Americans; a higher percentage of immigrants have college degrees than native-born citizens while, at the same time, a higher percentage of immigrants have not completed high school (Garcia & Cuéllar, 2006). The former hail predominantly from East and South Asia, while the latter include most of those who have immigrated from Mexico and Central America. Experience with schooling in their home countries would most certainly vary considerably between these two populations of ELLs, contributing to educational readiness upon arrival in the U.S. Narrative competence in particular may be

positively influenced by exposure to the types of literate activities children benefit from in preschool and school (Spinillo & Pinto, 1994).

Although Spanish-speaking ELLs share a common native language, it cannot be assumed to be the case that other important characteristics are shared. Spanish-speaking ELLs are a diverse population with respect to national origin, SES, immigration generational status, parental education, and experiences of schooling. However, most studies overlook these characteristics of their samples, focusing predominantly on language dominance and speech/language ability status when describing their participants (McFarland, 2011). This is consistent with prevailing paradigms of educational attainment, which attribute both successes and failures to individual ability and effort, overlooking important sociocultural and ecological variables. This oversight leads to an incomplete understanding of student learning, which contributes to patterns of the disproportionate representation of culturally and linguistically diverse students in special education (Harry & Klingner, 2006; Nasir & Hand, 2006; Ortiz, 1997; Seidl & Pugach, 2009). Adherence to the prevailing paradigm may limit our ability to develop a robust knowledge base on this population. The failure to include these important ecological and sociocultural variables both undermines our understanding of the population and limits the generalizability of findings with respect to the characteristics of their narratives.

Lack of Studies Investigating the Narrative Skills of ELLs with and without Learning Disabilities and the Relationship of those skills to Reading

Only two of the fifteen studies included in McFarland's (2011) synthesis specifically examined the relationship of ELLs' narrative skills to their reading achievement. Miller et al. (2006) was a large study with an ample sample size and importantly looked at cross-linguistic relationships at three grade levels spanning kindergarten and third grade. They did not include (or if they did include, they did not

report) students with learning disabilities to see if performance patterns differed for this group. Martinez-Roldán and Sayer (2006) looked at the role of language in bilingual children's reading comprehension. They elicited story retellings as data and used story grammar analyses as a measure of reading comprehension in Spanish and in English. Although they reported qualities of the children's narratives, they were not interested in narrative skill per se; rather, they were interested in the ways bilingual children drew upon their biliterate resources in both languages and in "Spanglish" to negotiate and to communicate the meaning of texts. Their sample included only four children in the third grade, all of whom were described in sufficient detail from a sociocultural perspective. While much can be learned from this case study regarding these bilingual children's narrative performance and reading comprehension, little can be generalized from it pertaining to the population of ELLs.

Incomparability of Findings between Studies due to Important Methodological Differences

A number of studies have compared the results of using different methods of eliciting narratives and analyzing narrative skills within a given sample of children and have found these different methods to produce widely differing results (Gazella & Stockman, 2003; Goldstein et al., 1993; Morris-Friehe & Sanger, 1992; Pearce 2003; Schneider, 1996; Schneider & Dubé, 2005; Shiro, 2003; Spinillo & Pinto, 1994). Some have found that the use of pictures as prompts, for example, may cause children to provide less information resulting in shorter narratives than they might produce under other conditions. The cause may be pragmatic; the storyteller may consider it unnecessary to include information that is readily available to both speaker and listener by way of the picture, which is in plain sight of both (Montague et al., 1990; Ripich and Griffith, 1988; Roth and Spekman, 1986). Spinillo and Pinto (1994) investigated

developmental changes in British and Italian children's narratives across three age groups and four storytelling tasks. The three age groups included four-, six-, and eight-year-olds, and the four tasks involved the following: 1) asking children to tell a story based on their own drawing; 2) asking children to tell a story based on a sequence of three picture cards placed in front of them; 3) asking children to simply make up a story; and 4) asking children to create a story and to dictate it to the experimenter, who would then share it with another child not present. Each child was given all four tasks and resulting narratives, which totaled 480 stories, were compared for sophistication of children's story schema as measured by a five-point holistic rating scale categorizing stories on a continuum from one (non-stories) to five (complete stories with a narrative structure that includes setting, characters, event(s), and resolution). Developmental differences were noted; the six- and eight-year-olds performed significantly better than the four-year-olds. Two hypotheses were tested with respect to the effects of experimental conditions on the children's narratives: 1) Stories produced by picture prompts (Tasks 1 and 2) would be less sophisticated than those produced without such prompts (Tasks 3 and 4); and 2) A higher level of narrative structure would result from the more structured tasks given similar conditions (e.g., stories from picture cards would be better than stories from drawings and dictated stories would be better than stories simply made up for no explicit purpose). Their first hypothesis was confirmed; elementary stories (rated 1-3) were more often the picture-elicited stories while more sophisticated stories (rated 4-5) occurred more frequently in the non-picture condition; this was true for all age groups. The second hypothesis was not confirmed, however.

The inability to directly compare results of studies is a major limitation of the current state of the research on the narrative skills of school age ELLs. Given the importance of narrative skills to the assessment of bilingual children, it would seem

especially crucial to identify a comprehensive measure that is sensitive enough to capture developmental as well as language-related differences in narrative skills while remaining clinically feasible, reliable, and efficient to use. This is an effort the present study seeks to undertake.

THE ASSESSMENT OF CHILDREN'S ORAL NARRATIVES

What are the salient features of a story? When a child reads or hears a story, be it anecdotal or fictional, what information or events do we expect him or her to predict, recall, or infer and to relate with other details in order to accurately get the overall meaning or gist of the story? In what ways does the organization of the story contribute to the comprehension of its meaning? Schema theory suggests that, as young children are exposed to stories (not just from books that are read to them, but also anecdotes and other forms of personal narratives they hear others tell), they begin to internalize a set of structural rules and components common to the stories they hear (Hughes et al., 1997; Mandler, 1988; Mandler & Johnson, 1977; Stein, 1988). They thus develop an underlying structure, or cognitive representation, of story structure that supports the comprehension and generation of oral stories (Hughes et al., 1997; Mandler & Johnson, 1977; Peterson & McCabe, 1983; Stein & Glenn, 1979). As children grow older and acquire literacy, schematic knowledge of stories aids in the comprehension (encoding and retrieval) of written text, specifically by facilitating both prediction and recall of information before, during, and after reading (Fitzgerald, 1984; Mandler & Johnson, 1977; Trabasso, Stein, & Johnson, 1981). Macrostructural analyses of narratives generally aim to uncover their structural characteristics in order to describe the overall organizational patterns of the narrative and levels of narrative development demonstrated by the storyteller (Hughes et al., 1997). Researchers of children's narratives have

documented progressive levels of episodic complexity and story completeness associated with children's age and developmental stage and there is general concordance between story grammar theorists regarding what kinds of stories typically developing children are generally able to tell at different ages (Applebee, 1978; Botvin & Sutton-Smith, 1977; Glenn & Stein, 1980; Hedberg & Westby, 1993; Hudson & Shapiro, 1991; Mandler & Johnson, 1977).

Scoring Systems Used in the Present Study

The present study includes three methods of narrative macroanalysis: Stein and Glenn's (1979) story grammar analysis, McCabe and Bliss' (2003) Narrative Assessment Protocol (referred to also as Narrative Assessment Profile in Bliss, McCabe, & Miranda, 1998), and the Narrative Scoring Scheme (Miller et al., 2006). Each method is described and a representative selection of literature involving each method is reviewed in the next session. Following that, research describing the oral narrative performance of SS-ELLs is reported.

Story Grammar Analysis

Story grammar analysis is a type of episodic analysis, a common approach to analyzing narrative macrostructure. In this approach, fictional stories are examined for the presence of story grammar elements and/or are assigned a certain level of story structure (Hughes et al., 1997). Story grammar refers to a cognitive system or schema for making sense of and retrieving information from stories (Mandler & Johnson, 1977; Stein, 1982). Stein and Glenn's (1979) approach, which is one of the most widely used with fictional stories (Hughes et al., 1997), posits a set of story parts or elements along with a set of rules governing the relationships between them. Their story grammar consists of a setting category plus an episode system consisting of seven story elements in

all: (1) Major and minor settings (in which the protagonist is introduced and the time and place of the story are established, respectively); (2) initiating events, which set in motion a problem situation or a state of affairs to which the protagonist must respond; (3) internal responses, including the emotional or affective response of the protagonist to the initiating event(s) and his or her subsequent goals or desires; (4) plans, indicating the protagonists strategies for obtaining his or her goals or desires; (5) attempts, or the protagonists goal-directed actions; (6) direct consequences related to the attainment or non-attainment of the goal; and (7) reactions, including how the characters respond to or are affected by the outcome. Stories are thus comprised of some or all of these elements. Major and minor setting elements provide essential contextual information, while the other six elements combine to form episodes, which establish temporal or causal connections between the elements. An episode must minimally include an initiating event or internal response, a goal-directed action or attempt, and a direct consequence related to the action or attempt.

In addition to quantifying story elements, story structure can be measured by assigning the story to one of several story structure levels (Glenn & Stein, 1980). These levels are typically represented on a scale from zero to seven, with zero indicating a sample that is unscorable (e.g., unintelligible, not told in the target language, etc.). Levels one and two are descriptive sequences that consist of unrelated objects, characters, or events, and are thus non-stories. Level three is a reactive sequence, in which events may be connected causally, temporally, or thematically, but there is no purposeful attempt to solve a problem and no goal-directed behavior. Level four is considered an abbreviated episode, in which there is an identifiable goal but no explicit planning or intentional action on the part of the protagonist to achieve the goal. Levels five and up constitute episodes, which may be incomplete, complete, or multiple (level 5), complex

(level 6), and up to embedded or interactive (level 7). See Appendix B.1 for a binary decision tree for determining story structure levels (Hughes et al., 1997). The first three story structure levels or sequences are typical of what a developmentally typical preschool-aged child will produce; level 4, the abbreviated episode, is typical of children around age 6; episodes (incomplete, complete, and multiple) are characteristic of the narratives of 7 to 8 year-old children and older, with complex, embedded, and interactive episodes not typically occurring until around 11 years of age (Hughes et al., 1997).

Much research employing story grammar measures has focused on describing and evaluating narratives produced by school-aged children with and without language and/or learning disabilities. These studies have produced mixed results, calling into question the ability of story grammar measures to detect significant differences between groups of students who are typically developing and those with disabilities (Hughes et al., 1997; Merritt & Liles, 1987; Ripich & Griffith, 1988; Roth & Spekman, 1986). Nevertheless, some consistencies in findings have emerged. For example, while monolingual children with language or learning disabilities are able to generate and retell stories that are structurally similar to those told by children without disabilities, their stories tend to contain fewer complete episodes, episodes of lesser complexity, and a lower frequency of story elements than the stories of their typically developing peers (Merritt & Liles, 1987; Roth & Spekman, 1986). The ability of story grammar analyses to differentiate the two groups appears to depend to some degree on the levels of analyses undertaken. Measures of global organization applied singly are less effective at predicting group membership than when they are combined or augmented with measures of microstructure such as grammaticality, cohesion, and syntactic complexity (Liles, Duffy, Merrit, & Purcell, 1995). The type of task used to elicit a story also affects the sensitivity of story grammar measures. Story retells tend to produce fewer between-group differences in story

grammar elements and episodic structure than story generation tasks because story structure is provided by the model story (Hughes et al., 1997; Schneider & Dube, 2005).

Roth and Spekman (1986) examined the spontaneously generated stories of children with and without learning disabilities over three age ranges using a story grammar approach. They analyzed the stories at the level of the proposition (for story grammar elements) as well as the episode (for episodic structure). Their participants were 48 students with LD and with 48 students who were normally achieving matched at the following age ranges: eight- to nine-years-old; ten- to eleven-years-old; and twelve- to thirteen-years-old. All subjects were native speakers of standard American English. Subjects with LD were included on the basis of not having language deficits that would interfere with their potential performance on the narrative tasks. Specifically, none were receiving remediation for oral language expression or comprehension in the areas of syntax, semantics, or phonology and they possessed sufficient skills in those areas to generate complete, grammatically correct, and meaningful sentences. Children were given the task to make up a story about something make-believe and were given unlimited time and a predetermined set of prompts and probes to complete the task. Children's stories were recorded, transcribed, and segmented into propositions (approximating a simple clause). Propositions were used as a measure of story length as well as a unit of meaning within a story. Segmented transcripts were analyzed using a modified version of Stein and Glenn's (1979) story schema. Each proposition was coded into one of the seven story grammar elements. Some modifications were made to those categories in order to accommodate dialogic instances (e.g., overt verbalizations of characters' responses) that occurred in the children's stories. Following the coding of propositions, stories were divided into episodes and episodic boundaries were marked. Episodes were then characterized by the story elements they contained and were

classified as being complete or incomplete using Stein and Glenn's criteria. Interepisodic relations (temporally sequential, temporally simultaneous, causal, and embedded) were also coded, as were the use of story markers (e.g., "once upon a time," "the end," etc.).

Several variables were measured. These included story length, number of episodes, episode integrity and structure, story category usage, interepisodic relations, the use of story markers, and the need for prompts. A two (group) by three (age) analysis of variance at the .05 level revealed significant differences between groups (F=17.09) and between ages (F=4.0) but no interaction effects between the two sources of variation. When compared with normally achieving students, students with LD: (a) produced significantly fewer propositions per story; (b) produced a significantly smaller proportion of complete episodes per total episodes; (c) included a significantly lower proportion of the category of Attempts in their incomplete episodes; (d) produced proportionally fewer episodes containing Responses, Attempts, and Plans. Regardless of age, the authors noted, "the learning-disabled subjects tended to omit the middle parts of a story, portions of which generally contain the cognitive planning, actions, and attitudes of the protagonist" (p. 14). There was only one significant main effect for age: the oldest subjects produced a significantly higher proportion of episodes with Setting statements.

The frequency of use of story elements (i.e., Setting, Initiating Events, Responses, Plans, Attempts, Direct Consequences, and Reactions) was measured and compared as well. Group differences were significant for Minor Setting statements (students with LD produced fewer of them) and for Initiating Events (students with LD produced proportionally more of them). As for interepisodic relations, students with LD produced proportionally fewer causal relations than normally achieving students and older students used more embedding linkages than younger age groups. Older students who were normally achieving used proportionally more concurrent episodes (the "And" category);

in fact, students with LD maintained a consistently low usage of And relations across all ages while their normally achieving peers demonstrated increasing use of And relations with age. Students with LD were therefore "less likely to connect episodes with the more complex temporal relations involving direct causality and simultaneity of events" (p. 15). There was no significant main effect for story markers, however for prompts, there was a main effect for age; older students were given fewer prompts.

In their discussion of findings, Roth and Spekman highlight that subjects with LD demonstrated a "relatively intact knowledge of story structure in that they used all category types in approximately the same order of saliency as their normally achieving peers" (p. 16). The order of saliency demonstrated by subjects with LD in this study was similar to that found in earlier gist recall studies: Attempts, Direct Consequences, Initiating Events, and Setting statements were the four most frequently included story elements. The most notable difference between normally achieving students and students with LD was in episodic integrity. Students with LD produced a significantly lower proportion of complete episodes than normally achieving students of the same age. In many ways, the authors concluded, the narrative performance of the students with LD resembled that of the considerably younger normally achieving students. Although children with LD appeared to possess sufficient narrative skills to produce at least some complete episodes, they demonstrated this awareness of narrative structure inconsistently, similar to developmentally younger children. Further, based on the types of omissions that were frequent in their episodes, students with LD appeared to demonstrate role-taking deficits, in which they failed to anticipate the comprehension needs of their listeners. They tended to omit the entire middle section of an episode, jumping from an initial introduction of a character and an initiating event to the outcome of the story. Similarly, they tended to omit information about a character's attempts, leaving the

listener to infer which actions connect the beginning of an episode to its resolution. Further supporting the observation that students with LD demonstrate developmentally lower narrative skills than their same-age normally achieving peers, the former tended to connect episodes in simpler and less mature ways that require less cognitive planning and organization. The authors concluded that, in contrast to previous reports, the story grammar approach was sufficiently sensitive to differences between students with LD and their normally achieving peers in their study. They attributed this outcome to the use of a story generation rather than a story recall task as well as the inclusion of episodic analysis and not just an analysis of individual story elements.

Montague, Maddux, and Dereshiwsky (1990) investigated the performance of students with and without LD on an oral story retell task as well as a written story completion task. Using Stein and Glenn's story grammar to analyze the narrative productions and using multivariate analyses of variance (MANOVA) to compare students grouped by ability status and age, their findings were similar to those of Roth and Spekman (1986). Students with LD were able to comprehend and to produce stories demonstrating an acquired knowledge of story schema, however their story productions differed significantly from those of their typically developing peers in terms of the amount and types of information included. In this study, subjects with LD recalled significantly fewer units of information and fewer internal responses of characters. The most salient differences in the writing task were found in the Internal Response, Direct Consequence, and Major Setting categories. Similar to Roth and Spekman, the authors concluded that students with LD have acquired a rudimentary but not fully developed story schema and that if their deficits in producing certain categories of information (e.g., internal response) were remediated, the total units of information in their stories would more closely resemble that of their typically developing peers. They observe, "if students

with learning disabilities could be taught to focus on the goals, motives, thoughts, and feelings of the characters in the stories they read and write, story length would increase proportionate to the increase in the internal response category" (p. 195). Like Roth and Spekman's sample, their subjects with LD demonstrated an ability to compose complete episodes both verbally and in writing; they, however, produced proportionally fewer complete episodes than the non-LD group.

Schneider & Dubé (2005) employed story grammar analyses to investigate the possibility that the amount of content included in children's narrative productions varies as a function of how story prompts are presented (orally, pictorially, or both). They included 44 typically developing children in kindergarten and 2nd grade as participants in their study. They presented each child with three different story stimuli. Each of the stories was centered on the same main character, a female hippopotamus, and each story included a different secondary character. Picture stimuli were taken from the same book and the oral versions of the stories included all story grammar elements, each of which occurred once in each story with the exception of the component, Reaction, which was included once for each character, thus twice in each story. Oral versions of stories were also controlled for story length in numbers of words as well as grammatical elements.

Each child participated in all three conditions: oral presentation without pictures, oral presentation with pictures, and pictures only. In each condition, the child was presented the story by one of the researchers and was asked to tell the story to a research assistant (a naïve listener). This was so the child could not assume that any prior knowledge of the story was shared and thus unnecessary to explicate. During the picture conditions, the listener was positioned behind a screen so the child would assume no shared access to the pictures. The mean number of story units generated under each

condition for each age group was calculated and analyzed using a two-way analysis of variance.

Findings revealed a main effect for grade level (F (2, 42) = 9.08; $p < .001$) as well as presentation mode (F (2, 84) = 27.53; $p < .001$). There was also a significant age by presentation interaction (F (2, 42) = 3.61; p = .031). Kindergarteners produced significantly more story grammar units in the combined presentation condition than in the pictures only condition; there were no significant differences for kindergarteners between pictures only or oral only stimuli or between the combined condition and the oral only condition. Children in second grade produced significantly more story grammar units during the oral only and combined conditions than they did in the picture condition but there were no differences between oral only and combined conditions for second graders. Out of ten possible story grammar units for each story, kindergarteners on average produced 5.82 in the pictures only condition, 6.55 in the oral only condition, and 7.09 in the combined condition. Second graders produced 6.32 in the pictures only condition, 8.32 in the oral only condition, and 8.64 in the combined condition.

The two groups of students performed similarly on the pictures only task, but differently when stimuli were presented orally. The authors considered several possible reasons for differences in performance under the different conditions and stressed the clinical implications of the findings: that care must be taken in choosing materials or elicitation methods for storytelling tasks and that story grammar methods of analysis may be sensitive to such variations. Others have reported similar findings and implications with respect to task effects on the length, content, and quality of stories children produce (Fiestas & Peña, 2004; Morris-Friehe & Sanger, 1992; Pearce, 2003; Schneider, 1996; Spinillo & Pinto, 1994).

Narrative Assessment Profile

Bliss, McCabe, and Miranda (1998) describe a multidimensional approach to analyzing children's narratives that was developed to evaluate discourse coherence and that would be capable of identifying specific aspects of a child's narration that are in need of intervention (McCabe & Bliss, 2003). They named their approach the Narrative Assessment Profile (NAP). The NAP evaluates the following dimensions of narration: topic maintenance, event sequencing, informativeness (which is multifaceted), referencing, conjunctive cohesion, and fluency. The first three dimensions represent more general or macrostructural aspects of narratives, whereas the referencing and conjunctive cohesion represent more specific, or local narrative discourse functions. Fluency is included as well, as it pertains to the manner of production, and dysfluency is a common trait of the narratives of children and adults who are language impaired (McCabe & Bliss, 2003; Peterson & McCabe, 1983). As a lifespan approach, the NAP is applicable to evaluating the narration of children and adults, with and without language impairment. It is a clinically useful assessment in that it provides a profile of relative strengths and weaknesses across a variety of narrative discourse dimensions and can be used to plan and monitor intervention. Furthermore, it is flexible and useful in the evaluation of discourse impairments associated with a variety of disabling conditions, including specific language impairment (SLI), autism, brain injury, hearing impairment, and intellectual disabilities. Finally, it is argued to be a more culturally sensitive measure of narrative discourse than story grammar analyses in that it avoids the latter's bias in favor of distinctively western European story structures (McCabe & Bliss, 2003). The six dimensions of the NAP are described next, followed by a description of some of the research that has been conducted using this instrument.

All six dimensions of the NAP are rated on a three-point scale (0-2), to describe behavior that is "appropriate", "variable", or "inappropriate". A designation of appropriate (given 2 points) signifies that the behavior occurs frequently enough to promote and maintain discourse coherence. Variable behavior (given 1 point) in a given dimension indicates that the level of performance occasionally reduces discourse coherence but that the narrator does demonstrate some strengths in that dimension. A behavior is considered inappropriate (and is given 0 points) when its frequency diminishes or compromises discourse coherence. The NAP does allow scorers to suspend judgment and to indicate that a particular dimension "needs further study," as well.

Topic maintenance refers to how well all the various utterances in a narrative relate to a central theme. In order to achieve appropriate topic maintenance, narrators must avoid digressions, including irrelevant, tangential, or vague utterances that detract from the theme, disrupting discourse coherence. The ability to maintain topical coherence emerges in preschool and is developed during the school years. In their research, the authors have found that children with SLI tend to deviate from the topics of their narratives.

Event sequencing involves the order of presentation of events. Events must be presented in either chronological or logical order and should correspond with real-life ordering of events unless the narrator explicitly indicates that he or she will violate the expected order. The narrative pattern described by high point analysis as "leapfrogging" is an example of disjointed event sequencing where events are presented out of order and/or where critical events are omitted (McCabe & Rollins, 1994). Because the listener cannot keep track of events presented in this way, the coherence of the narrative is

compromised. Leapfrogging narrative patterns are characteristic of the narratives of children under the age of five as well as some children with language disorders.

Informativeness refers to the elaboration necessary to make a story complete. The NAP evaluates three kinds of informativeness: factual information of the type and level of detail required by a police officer; the embellishment necessary to make a narrative engaging to listeners, as would be requested of a teacher; and the narrative "ingredients" which constitute the recipe for a good narrative (Labov, 1972), specifically: description, action, and evaluation. By scoring each of these facets of informativeness, the NAP weighs this dimension heavily in the overall evaluation of narrative strengths and provides a means of identifying specific areas of deficit. The omission of information necessary to well-formed episodes according to other story schema (e.g., Stein and Glenn's story grammar elements) is captured under this category as well as evaluative features. In this way, the NAP appears to combine the best features of both episodic (story grammar) analysis and high point analysis. When school-aged children with normal language omit information in narratives, it is usually information of a kind that can be easily retrieved by context and inferred by the listener. Children with language and learning disabilities, on the other hand, tend to omit crucial information that cannot otherwise be inferred, a pattern that may reflect a limited awareness of the listener or audience and a limited ability to identify with and anticipate a listener's comprehension needs.

Referencing involves the adequate identification of people and things within a story. Poor referencing confuses listeners because pronouns have ambiguous or no antecedents, because nouns are used where pronouns would be expected, or because erroneous pronouns are used. There is some documented variation in the development of referencing in children of different socioeconomic statuses. Hemphill (1989) noted that

low-income children and adults tended to use more unspecified pronouns, perhaps reflecting a cultural norm whereby the listener is expected to collaborate in the construction of the discourse. By contrast, middle class speakers were far more explicit, providing all necessary information for the listener to make sense of a narrative. Middle class children demonstrate referential adequacy beginning at about three years of age. Children with language impairments often have inadequate referencing abilities. They tend to use more nonspecific pronouns or demonstrative pronouns (such as "this" or "that") where personal pronouns would be more appropriate.

Conjunctive cohesion refers to words or phrases that link utterances (i.e., and, then, but, because, when, so) and thus are essential to the ability of the listener to discern relationships between utterances in a narrative. Cohesion includes the following semantic links: coordination (how a series of events are described), temporal (events related in a time sequence), causality (events tied together by cause and effect relationships), enabling (events that establish preconditions for another event), and disjuncture (semantic contrasts between two clauses). Pragmatic links are also aspects of cohesion, and include cohesive devices signifying beginnings, endings, changes of focus, and explicitly stated chronology violations, all of which serve to enhance discourse coherence. The ability to use cohesive devices develops throughout the elementary grades and even children as young as four are able to use conjunctions for semantic as well as pragmatic functions. Children with SLI often use conjunctions inappropriately (for example, committing semantic violation by reversing cause and effect relationships through the misplacement of the conjunction, "because").

Fluency refers to uninterrupted discourse. Common sources of disruption are false starts, corrections, and unnecessary repetitions. Dysfluencies are common in the speech of two- to four-year-old children and decrease thereafter. Some children with

language disorders continue to exhibit dysfluencies in their discourse into their school years.

Miranda, McCabe, and Bliss (1998) sought to describe the discourse coherence of a sample of children with SLI using five of the six dimensions of the NAP as variables. In this analysis, which predated the development of the NAP as an instrument, the dimension of explicitness was used to detect referential abilities and informativeness was not included as a variable. They examined three groups of children: a group of children with SLI and two comparison groups – one matched on age and the other matched on language maturity. All children had normal abilities and intelligence outside of specific language impairment. Ten males aged eight to nine comprised the SLI group. The comparison group matched on age consisted of ten boys without SLI who were of the same age, ethnic, and socioeconomic background. The comparison group matched on language level consisted of ten typically developing boys ages five to six whose scores on the Index of Productive Syntax matched those of the boys with SLI.

Personal narratives were elicited of all subjects using the Conversational Map procedure. Five verbal prompts were presented to each child in random order. To prepare for analysis, the recorded narrative samples were transcribed and segmented into propositions as the unit of analysis. Hierarchical relationships between propositions were then displayed in an outline. Finally, dysfluencies were identified. Topic maintenance was assessed by identifying thematic propositions, or those propositions related to one experience. Nonthematic propositions were also identified and coded into two types: script-like segments and miscellaneous segments. Event sequencing was assessed by coding for ordered and disordered segments. Ordered discourse patterns were coded as being either single events or multiple events, and disordered patterns were coded as a leapfrogging narrative. To assess explicitness, topical narrative propositions were coded

as being either explicitly stated or implicit (implied by the child). Four types of errors of implicit reference were included in this category. The extent to which narrators were sufficiently explicit was measured by the number of propositions each rater independently identified as implicit. Conjunctive cohesion was analyzed by identifying the semantic and pragmatic functions of conjunctions as well as errors in the usage of conjunctions. If a conjunction did not serve a semantic purpose, a pragmatic analysis was performed. Agreements between raters as to whether conjunctions were used semantically, pragmatically, or in error were calculated. Fluency was analyzed by calculating the frequency of dysfluencies, specifically reformulations, repetitions, and discontinuations, or false starts.

Results revealed significant group differences between the SLI group and one or both of the comparison groups in each of the dimensions. Children with SLI: (a) produced more off-topic utterances; (b) engaged in more leapfrogging; and (c) were much less explicit, placing considerably more burden on their listeners to make sense of the greater proportion of implicit references and other crucial information. Importantly, the analysis was also able to identify areas in which the children with SLI displayed unexpected strengths. These areas included the use of connective devices and fluency. Children with SLI produced more connectives than the matched language ability group, but fewer than the comparison group matched by age. However, children with SLI also committed fewer errors with connectives than what was expected based on the results of other studies. The authors attributed this to the method of analysis, which looked at the pragmatic functions of connectives and not just their semantic functions, so that what may have been counted as semantic error in other studies, was credited in this study as having a pragmatic function. With respect to fluency, children with SLI did not differ from the other groups in the ways that were expected (greater numbers of repetitions,

reformulations, and discontinuations). However, they did demonstrate a higher ratio of total dysfluencies when compared with their same-aged peers, mostly due to a greater frequency of reformulations, which result from difficulties in word retrieval, grammatical repair, and attempts at formulating sentences. The authors concluded that the overall dimensions assessed in the study provided a profile of narrative discourse abilities, including the relative strengths and weaknesses of the narrative discourse of children with SLI.

The NAP was developed and described in later work (Bliss et al., 1998; McCabe & Bliss, 2003), in which the authors demonstrated its clinical usefulness in assessing the strengths, weaknesses, and intervention needs of children and adults of various abilities and cultural and linguistic backgrounds.

Narrative Scoring Scheme

John Heilmann and his colleagues at the Language Analysis Lab at the University of Wisconsin-Madison developed a measure of children's overall narrative organization skills that was designed to be sensitive to a wider range of ages and sampling contexts (i.e., the varied levels of support inherent in different task conditions such as story retell versus story generation). To develop their instrument they reviewed literature to identify the features of more advanced narrative productions as well as scoring methods, beyond story grammar analyses, that would be sensitive enough to evaluate these more advanced features. Thus, their Narrative Scoring Scheme (NSS) incorporates basic story grammar features *plus* "specific *types* of language that define a literate style of speaking" (Heilmann, Miller, & Nockerts, 2010, p. 609). A literate style of speaking is characterized by abstract language such as metacognitive verbs, used to describe characters' thoughts and mental states, and metalinguistic verbs, used to describe

characters' speech and dialogue. It also includes the use of cohesive devices, including referential cohesion, conjunctive cohesion, and lexical cohesion. The developers of the NSS attempted to increase the sensitivity of the measure by relying on a holistic rating process using a sufficiently broad scale. The NSS thus includes seven aspects of narrative organization for which examiners must make qualitative judgments on a five-point scale, resulting in 35 possible points per narrative. The seven sections of the NSS are: introduction, conflict resolution, and conclusion (modeled after story grammar proposals); mental states and character development (measuring use of literate language); and referencing and cohesion. In each section, a score of 1 indicates immature performance, 3 signifies emerging skills, and 5 indicates proficiency.

Heilmann and colleagues (2010) sought to compare the measurement properties of the NSS with other methods of evaluating children's narratives, including a plot and theme analysis of key story grammar elements and two holistic rating measures of narrative organization using ordinal scales: Applebee's story structure levels, and Stein's scoring scheme. The authors hypothesized that the NSS would be more developmentally appropriate for their sample of five- to seven-year-old children who produced narratives using the story retell procedure. Once narrative samples were collected and analyzed using each scoring system, the distribution of scores was evaluated. Specifically, the investigators were interested in determining whether the distribution of scores from the NSS was less skewed than scores from the three other measures of narrative organization. Narrative samples were collected from 129 typically developing, English-speaking children in Southern California. To elicit narrative samples, examiners read a target story to the children, who followed along in a wordless picture book (*Frog, Where are You*?; Mayer, 1969). The children were then asked to retell the story to the examiner.

Narrative retells were recorded and transcribed and then subject to scoring by each of the measures, including the NSS.

Upon subjecting the 129 story samples to each of the four narrative organization measures, the authors produced histograms depicting the distribution of scores for each measure. Each of the three methods that were compared with the NSS demonstrated ceiling effects: their distributions of scores clustered at the top of the scale range of each measure. The NSS, on the other hand, resulted in a fairly normal distribution, with most scores centered on the mean (20.1 on the 35 point scale) and no scores at either extreme of the scale (sample range = 11 to 26). The NSS was indeed less skewed than the other scoring systems. Additionally, the NSS was sensitive enough to reveal differences between three narrative productions specifically chosen to reflect a range of performance (poor, average, and best). Whereas the other three measures were unable to distinguish between the average and the best narrative productions, the NSS was able to distinguish all three.

The NSS was designed to be clinically efficient yet sensitive enough to detect narrative skill growth in older children while distinguishing typical development from language learning difficulties in younger children. The developers of this instrument have compiled several databases that serve as referents for both monolingual English speakers and Spanish-speaking English language learners. They attribute the sensitivity of their measure to its incorporation of children's use of literate language and cohesion, which are later developing skills. Also, by using holistic examiner judgment the NSS is able to detect the perceptual aspects of narratives that discrete scoring schemes are unable to detect.

CHARACTERISTICS OF THE NARRATIVES OF SPANISH-SPEAKING ENGLISH LANGUAGE LEARNERS

McFarland (2011) completed a systematic literature review, the purpose of which was to describe the characteristics of both the Spanish and English narratives of SS-ELLs of different ages. Included in this review were studies (see Table 2.1) that analyzed narrative microstructure (language productivity) and/or narrative macrostructure (variations of story grammar analyses) of the oral stories of SS-ELLs produced in English and in Spanish. Relevant literature was located systematically using keyword searches and exhaustive searches of specific journals to identify relevant research articles published in peer-reviewed journals between 1990 and 2010. To be included, articles had to be based on empirical research and had to report the characteristics of narrative samples elicited of SS-ELLs between the ages of 4 and 11 (or prekindergarten through 6th grade). Articles were read and coded systematically to facilitate aggregation and description of the collective results of comparable measures. Findings of the investigation are reported in the section that follows.

Table 2.1

Studies Examining Spanish-English Bilingual Children's Oral Narratives

Authors/Year	N	Mean Age	Disabilities
Bedore, Peña, Gillam, & Ho (2010)	170	5;6	LI
Fiestas & Peña (2004)	12	4;0 to 6;11	None
Gutierrez-Clellen (1998)	57	7;9	None
Gutierrez-Clellen & Hofstetter (1994)	77	5;1 6;6 8;6	None
Gutierrez-Clellen & Iglesias (1992)	46	4;0 6;0 8;0	None
Gutierrez-Clellen (2002)	33	7;3 to 8;7	None
Gutiérrez-Clellen, Simon-Cereijido, & Wagner (2008)	71	5;7	47 TD 24 LI
Martinez-Roldán & Sayer (2006)	4	3rd grade	None reported
Miller, Heilmann, Nockerts, Iglesias, Fabiano, & Francis (2006)	1531	K-3	None reported
Montanari (2004)	3	5;4-5;8	None reported
Muñoz, Gillam, Peña, & Gulley-Faehnle (2003)	24	4;4 5;4	None
Schoenbrodt, Kerins, & Gesell (2003)	12	6 to 11	None reported
Simon-Cereijido & Gutierrez-Clellen (2009)	196	5;7	126 TD 70 Lang Delay
Uccelli & Paez (2007)	24	5;6-6;6	None
Uchikoshi (2005)	108	5;7	None reported

Note: TD = typically developing; LI = language impairment; Lang Delay = language delay; None = no children with disabilities were included in sample; None reported = disability status of participants was not reported.

Description of Studies Included in the Review

Fifteen studies met inclusion criteria and were included in the analysis. Two of the fifteen studies (Schoenbrodt, Kerins, & Gesell, 2003; Uchikoshi, 2005) investigated the effects of narrative interventions on the narrative skills of bilingual children while the other thirteen studies primarily described bilingual children's narrative skills. The fifteen studies in aggregate examined the narratives of a total of 2,368 Spanish-speaking English learners between the ages of four and eleven. Gender was reported for 401 (19%) of participants in aggregate. Of those for whom gender was reported, 55% were boys and 45% were girls. Language dominance was reported for 99% (n=2,344) of the participants. Of these, 90% (n=2,114) were Spanish dominant (used Spanish at least 60% of the time), 4% (n=102) were English dominant (used English at least 60% of the time), and 6% (n=128) were considered "balanced bilingual," which is defined as using either Spanish or English 40–60% of the time, and the other language the remainder. Ability status was reported for 710 or 30% of participants. Of these, 264 (37%) participants were reported to have disabilities; all of these students were reported as having speech/language impairments.

Socioeconomic status was reported for 359 (15%) of participants, all of whom qualified for their school's free or reduced lunch programs. Of participants for whom national origin or descent was reported (n=439 or 19%), over 310 (71%) were of Mexican descent, 93 (21%) were of Puerto Rican descent, and 36 (8%) were from other Latin American countries. All studies reported participant's ages; however, for 69 (2%) participants, only an age or grade range was given (ages 6-11, grades K-5). Of the remaining 2,323 (98%) participants, 60 (3%) were age 4 or in Pre-K, 949 (40%) were age 5 or in Kindergarten, 459 (19%) were age 6 or in 1st grade, 388 (16%) were age 7 or in

2nd grade, and 467 (20%) were age 8 or in 3rd grade. Six of the fifteen studies reported findings disaggregated by age.

Methods for establishing language dominance were reported for the majority of the studies. In nearly all cases, parent and teacher reports of home and classroom language use combined with more formal measures of language proficiency collected in each language were used to determine language dominance. Information regarding parent education and the instructional programming of participants was reported in fewer than half of the studies.

Language Productivity Characteristics of ELLs' Narratives

Twelve of the fifteen studies applied measures of language productivity to narrative analysis. The most commonly used measures included in descending order, Mean Length of Utterance (MLU), Number of Utterances (NU), Number of Different Words (NDW), Percent of Grammatical Utterances (%GR), Total Number of Words (TNW), and Subordination Index (SI), a measure of sentence complexity.

The number of different words children used to tell their stories generally increased with age and ability. Similarly, total number of words was greater for older children in the studies that disaggregated data by age group. This performance pattern was consistent for each of the other language production measures as well. Age, ability, and language dominance were generally associated with longer, more complex, and more grammatical utterances and wider vocabulary usage. One exception was Miller et al.'s (2006) study of 1,531 Spanish-dominant children whose mean length of utterance did increase with age for both languages, but whose English utterances were, on average, longer than their Spanish utterances. Their procedure for segmenting ellipted clauses into separate utterances could explain this pattern. In ellipted clauses, a single subject is

associated with multiple predicates (e.g., the boy ran, jumped and fell). The use of ellipted clauses is more prevalent in Spanish, a language in which information about the subject is encoded in the verb (e.g., el niño corrió, saltó, y se cayó) where the verbs correr, saltar and caer are conjugated in the 3rd person preterite; had el niño (the boy) already been mentioned in a previous sentence, there would be no need to restate this noun phrase or a pronoun at the beginning of the sentence (e.g., El niño fue al parque. Corrió, saltó, y se cayó; The boy went to the park. He ran, jumped, and fell). Thus, while MLU increases over time for both English-speaking and Spanish-speaking children, the trajectory is flatter for Spanish-speaking children (Bedore & Peña, 2008; Bedore, Peña, Gillam, & Ho, 2010).

Across languages, results reported for narratives elicited in the dominant language generally displayed a tighter distribution; standard deviations tended to be smaller for all measures in the dominant language while greater variation was evident in the weaker language.

Story Grammar Characteristics of SS-ELLs' Narratives

Ten of the fifteen studies analyzed children's narratives using some form of story grammar analysis. Seven of the ten applied analyses consistent with either a plot and theme analysis or a holistic analysis; because they used different analyses, and/or because of idiosyncrasies in the ways results were reported, the other three could not be compared to those seven studies. As was the case with the language production measures, story grammar scores consistently increased with age. Furthermore, higher story grammar scores were associated with the dominant language (Miller et al., 2006). Another pattern emerged for the studies that measured story grammar in both languages at different times within the same subjects (Montanari, 2004; Schoenbrodt, Kerins, & Gessell, 2003;

Uccelli & Páez, 2007). While story grammar scores improved over time for each language, the improvement was greater in English regardless of language dominance. This was true of both the intervention and non-intervention studies that measured children across time and was attributed in part to the children gaining linguistic resources in their L2 as a result of exposure to English (Montanari, 2004). None of the studies that included story grammar analyses included samples with any reported disabilities, so performance level by ability status could not be compared.

Both intervention studies demonstrated that the story grammar of typically developing ELLs could be improved through narrative intervention conducted in either language. The interventions were designed to increase narrative skills by teaching children to recognize and attend to the components and structure of stories, such as characters, setting, and plot (as in Schoenbrodt et al., 2003) or by simply exposing children regularly to stories with consistent narrative structure (as in Uchikoshi, 2005).

Schoenbrodt et al (2003) conducted an 8-week structured narrative intervention in which children were taught first to use and then to create a story grammar marker (SGM) to aid their comprehension of narrative events and styles in stories. The SGM is a tangible marker using symbols to represent story grammar components, including main character, setting, events, conclusion and the main character's internal response at different points in the story. The researchers found that their Spanish-dominant subjects who received the intervention in their native language performed significantly better on post-measures of narrative style (including grammaticality, cohesion, and pragmatic features of the child's storytelling) than did the control group who received the same intervention but in English. Story grammar measures improved significantly from pre- to post- intervention for both conditions and there were no significant between group

differences. Neither group showed increased production of language as measured by number of utterances and total number of words.

Uchikoshi (2005) also found positive effects of a narrative skills intervention delivered in English for Spanish-speaking ELLs in kindergarten. Bilingual kindergarteners in Uchikoshi's study were assigned to one of two groups, both of which viewed a 30-minute book-based educational television program during school hours three times per week for a total of 54 episodes. The experimental group watched Arthur, a program in which two stories, each with a plot including a conflict and resolution, were presented each episode. The control group viewed Between the Lions, a program that introduced a book each episode but which emphasized discrete literacy skills such as phonological awareness, the alphabetic principle, phonics, punctuation, and the conventions of written English.

The study tested the hypothesis that bilingual students exposed regularly to Arthur with its emphasis on narrative structure would develop stronger narrative skills than classmates exposed to a similar book-based television program, but one that de-emphasized overall narrative structure in favor of an approach in which parts of the text are highlighted. The two groups were compared on the rate and level of change from pre- to post- intervention on a combined narrative measure in which children's stories were coded for story structure, events, evaluation, temporality and reference, and storybook language; they were also compared on story length and sentence complexity as measured by TNW and MLU, respectively. Correlations between measures at both times and with initial Spanish and English vocabulary were reported, as were means and standard deviations for all children and for children by experimental group and gender.

Mean differences for all measures were greater for the children who viewed Arthur. Individual growth modeling showed steeper growth trajectories on the combined

narrative measure for students in the experimental group, but also revealed much variation in growth trajectories for individual children. After adding various predictor variables to the model, gender and initial English vocabulary had significant effects on initial levels of the combined narrative measure (CNM). Boys and children who started kindergarten with higher levels of English vocabulary scored higher on the initial CNM, however, neither variable was associated with rate of growth.

The fifteen studies examined in the systematic literature provide a snapshot of the state of research on the oral narratives of SS-ELLs. First, sample descriptions ranged from sparse to robust. Providing ample information about the backgrounds of participants is crucial to the generalizability of any study of SS-ELLs, given their heterogeneity in the U.S. population. In terms of language productivity, or the surface, microstructural aspects of oral narrative performance, performance on various measures generally increased or improved with age and with ability. This is a predictable finding consonant with what is known about monolingual populations. The difference was that, when compared with Spanish performance, English performance was more variable exhibiting a greater range. With respect to macrostructural aspects of narrative performance as measured by story grammar analyses, the narrative organization skills demonstrated by the samples also increased with age and higher quality stories were associated with the dominant language. Average story grammar holistic scores (converted to percentages) across studies were incomparable due to differences in elicitation conditions. Intervention studies demonstrated that oral narrative skills could be improved through explicit instruction and exposure to narrative texts. In both of these studies, the greatest improvement in scores occurred with narrative samples collected in English.

SUMMARY

Empirical research on the English oral narrative skills of SS-ELLs suggests that it is difficult to define what constitutes "typical" performance at any given age for this population. Performance patterns reflecting high variability may be related to several factors, including level of language proficiency in English and development in the native language, exposure to former schooling, narrative task effects, and differences in analysis procedures, making it difficult to interpret narrative performance. This underscores the importance of the current study's goal, which is to describe the characteristics of a sample of SS-ELLs' English oral fictional narratives and scrutinize the differences that result from applying three different scoring systems to their analyses. This study's findings will facilitate better interpretation of SS-ELLs' English oral narrative performances and will inform the design and development of appropriate systems for evaluating them. The information generated by well-designed systems may provide teachers and related service professionals important knowledge of ELLs' narrative organization skills, contributing to a more complete profile of an individual ELL's literacy-related English oral language development. Given the population's historically poor academic outcomes and patterns of disproportionate representation in special education, it is hoped that this study's contribution to research and practice will result in better-informed professionals and improved services, enhanced opportunities to learn, and ultimately, better academic outcomes.

CHAPTER THREE

Method

RESEARCH QUESTIONS

Researchers and practitioners interested in assessing the narrative skills of ELLs have a number of methods at their disposal (Hughes et al., 1997; Peterson & McCabe, 1983). However, to facilitate interpretations of student performance, what is needed is a better understanding of the characteristics of ELLs' oral narratives and the ways in which different systems for narrative analysis generate useful information about a child's narrative language skills.

This is an ex-post facto, exploratory study based on a subset of data collected for The University of Texas at Austin model demonstration project, *Determining Special Education Eligibility for the Bilingual Exceptional Student: Early Intervention, Referral & Assessment (BESt ERA)* (ED 524B, 2006-2010), funded by the U.S. Office of Special Education and Rehabilitative Services (OSERS) and the Texas Education Agency (TEA).

The purpose of the study is fourfold: (a) to describe the characteristics of the English stories of Spanish-speaking ELLs; (b) to compare how each of three scoring systems characterizes the sample according to its own criteria; (c) to identify the stable features of narratives that are rated consistently across scoring systems; and (d) to identify criteria for a high quality narrative scoring system for evaluating the English oral narrative skills of Spanish-speaking ELLs. The following research questions guide the study:

1. What are the characteristics of second grade Spanish-speaking ELLs' stories, orally narrated in English, using the following methods of analyses?

a. Story grammar analysis
b. Narrative Assessment Profile
c. Narrative Scoring Scheme

2. How does each scoring system characterize the sample in terms of expected narrative performance according to its own criteria?
3. What are the distinguishing features of narratives whose scores are consistent (e.g., high, average, and low) across measures?
4. What features must a narrative scoring system have in order to provide teachers with quality information that will help them design instruction and interventions for SS-ELLs?

CONTEXT FOR THE STUDY

Model Demonstration Project

The purpose of the BESt ERA model demonstration project was to design a professional development and technical assistance model to address the disproportionate representation of ELLs in special education. Working in partnership with a Central Texas Independent School District (ISD), the project was implemented over four years at two of the ISD's bilingual elementary school campuses. Two additional schools served as comparison sites. Project staff worked with campus administrators and teachers to put into place a prereferral process for struggling ELLs. Project results informed the development of professional development modules, which were disseminated via a training-of-trainers conference. The BESt ERA model promotes early intervention with ELLs who are struggling with literacy; therefore the Model Demo project supported an after-school reading and ESL tutoring program at one of the participating campuses. Narrative samples were collected as part of that program in order to help facilitate an

understanding of the relationship between academic language and literacy development and to provide information which would support campus-based problem-solving processes.

Participating Sites

The participating school district is a large urban district with a current annual enrollment of approximately 87,000 students attending 124 school campuses. District data for the 2007-2008 school year indicate that 58% of the district's students were Hispanic, 60.8% were economically disadvantaged, 28.3% were considered Limited English Proficient (LEP), and 57% were at-risk. The data used in the current study come from student participants in the model demonstration project who were enrolled in a bilingual education program at an elementary school serving children in grades preK through 5. The participating campus had enrollment of over 80% Hispanic students and over 60% LEP students.

TELL ME A STORY: SCORING AND ANALYSIS OF ENGLISH ORAL NARRATIVE SKILLS OF SECOND GRADE SPANISH-SPEAKING ENGLISH LANGUAGE LEARNERS

In the sections that follow, the methodology of the current study is described.

Research Approval

An IRB protocol (Number 2013-03-0074) was submitted with the Office of Research Support (ORS). Upon review, they determined that the current study did not meet the requirements for human subject research as defined in the Common Rule (45 CFR 46) or FDA Regulations (21 CFR 50 & 56) and therefore IRB review and oversight was not required.

Participants

Participants (n=42) for the current study came from the Model Demo campus at which the project-supported after-school tutoring program was implemented during the spring semester of 2008. All 42 participants were in the second grade. Of these, 59.5% (n=25) were female and 40.5% (n=17) were male. All were classified as limited English proficient and were participating in a transitional bilingual education program in which they received native language literacy instruction as well as English as a second language instruction.

Language Proficiency Level of Participants

The Language Assessment Scales-Oral (LAS-O) – English (De Avila & Duncan, 1990) is a screening device that measures the oral language skills "necessary to succeed in an American mainstream academic environment" (Clearinghouse on Assessment and Evaluation, 2012). It assesses the phonemic, lexical, syntactical, and pragmatic language skills of English language learners, grades 1-12. The LAS-O - English is a standardized measure, which was normed on 3,600 students in Texas and California. It is designed to: 1) aid in the identification of students with limited English proficiency; 2) help determine language dominance; 3) identify placement needs; and 4) determine proficiency levels. It is also intended for use as a measure of change over time. The Story Retelling section utilizes a procedure that is an adaptation of what is known as "focused holistic scoring." For the procedure to remain reliable and valid, scorers must be proficient, literate speakers of English and participate in a reliability exercise, attaining a reliability level of 90%.

The LAS-O uses a scale of 0 to 5 to evaluate the story retells that children complete. The task involves listening to a recorded story while the examiner points to the corresponding illustrations provided in a four-picture sequence. The child is then asked

to look at the pictures and retell the story that they just heard. Their responses are audio recorded and transcribed. Transcripts are then verified and scored. A score of 3 describes a response that includes a recognizable story line but contains errors in grammar, syntax, vocabulary, or usage that would be uncharacteristic of proficient speakers of standard American English. Scores of 4 and 5 are given to stories that are complete, fluent, and increasingly articulate, and well elaborated. Where language errors surface, they are not uncharacteristic of proficient speakers of standard American English, nor do they detract from the story line.

The LAS-O story retell task was administered, transcribed and scored by members of the Model Demo research team. Scoring disagreements were resolved through consensus. Results for the 82 LAS-O story retells completed by the 42 participating students in this study averaged 2.6 with a median score of 3 (see Table 3.1). Most students were able to communicate the story's basic storyline but exhibited notable errors and dysfluencies that would be unlikely to be made by proficient speakers of American English.

Table 3.1

Mean and Median LAS-O Scores by Testing Session

	N	Mean	Median	SD
LAS-O Mar	42	2.57	3	0.67
LAS-O May	40	2.7	3	0.61
TOTAL	82	2.64	3	0.64

Data Sources

Oral narratives elicited by the Tell A Story about a Picture (TASP) task (described below) and transcribed by Model Demo research staff are the primary data

source for this study. Forty-one of the 42 students completed the TASP assessment twice during the spring semester, 2008: at the start of the after-school tutoring program in March, and again at the conclusion of the program in May. One student completed the TASP assessment only once, during the first administration in March. As a result, 83 English TASP transcripts are included in this study. Of the 41 subjects who told two stories, 36 told a story about the same picture both times and 5 told one story about each picture. All 42 participants completed the English LAS-O story retell at the first administration in March, and 40 participated in the second administration in May, resulting in 82 English LAS-O scores. The LAS-O scores are included in this study in order to help describe participants in terms of their levels of oral story retelling skills and to provide a norm-referenced measure with which to compare this study's results.

Instruments

Telling a Story about a Picture (TASP). The TASP, based on the Oral Language Evaluation (Silvaroli, Skinner, & Maynes, 1977), is a criterion-referenced, standardized assessment in which students are asked to generate a story using picture stimuli. The pictures used as stimuli must depict a topic familiar to children with enough activity to elicit story elements, which include setting, an initiating event or problematic situation, some attempt to resolve the problem, a consequence, and inference, prediction, or evaluation (Westby, 1992). To elicit this sample, two pictures were used to prompt stories (see Appendix A). One depicted a circus scene where a lion appears to have just escaped from his cage. A boy is running from the lion while onlookers, including a clown and the lion tamer, are watching astonished. The second picture shows a street in what appears to be an urban neighborhood where two or three boys are playing baseball. A window in a building on one side of the street is broken and a woman is standing

beside it, angrily pointing toward the boys on the street. A police officer has grabbed the wrist of one of the boys, who has dropped the baseball bat while running.

Students were given the instructions: "Tell me a story about this picture. Tell me the very best story you can tell me." Their responses were audiotaped, transcribed and scored to determine the story level on a scale of 0 to 5, where 0 signifies a story that cannot be coded (e.g., is unintelligible); 1 is a non-story in which a child predominantly labels objects in the picture; 2 is a non-story in which a child lists or describes actions or events depicted; 3 is an incomplete story in which a child conveys causal relationships between actions or events and conveys a main idea; 4 is a complete story, which necessarily includes an initiating event or problem, an attempt to resolve the problem, and a consequence or resolution; and 5 is a complete story with mood, evaluation, or inference. Model Demo research team members administered, transcribed and scored the TASP. The story transcripts that resulted from TASP administration are the primary data sources of this study, however the TASP scores from the original study were not used.

Confidentiality of Data

The Model Demonstration data are maintained in secure files in the UT Austin Office of Bilingual Education. Student participants were assigned identification numbers. Data used in the current study list only these assigned ID numbers; no names or other identifying information, including names of school campuses, are maintained in this study's records.

NARRATIVE ANALYSIS

Preparation of Story Transcripts

Prior to analysis, each transcript was segmented and coded for microanalysis (language productivity measures) using the coding conventions of the Systematic

Analysis of Language Transcripts (SALT) software (Miller, Andriacchi & Nockerts, 2011). Transcripts were segmented line by line into C-units, which are defined as an independent (or main) clause with its modifiers (Loban, 1976). Subordinate clauses were not segmented separately; rather each remained with its main clause and was counted as a single utterance. Co-ordinate clauses were separated, including those that shared a single explicit subject, in which case the fragmented utterance was marked with an "[F]" to indicate an allowable subject omission. According to SALT conventions, specific characters were used to mark the ends of complete and abandoned utterances. Both child speech and key examiner speech (including examiner prompts, interruptions, insertions and questions) were coded. Transcript information, including narrative identification information, identification of the picture prompt (baseball or circus), and any additional transcription comments were recorded as a preface to each transcript. Transcripts were coded to mark the following variables:

1. Prompt Drivenness: A categorical judgment of whether the narrative adheres to the information provided by the picture fully (PD), partially (PPD), or not at all (NPD).
2. Story Interpretation: A categorical judgment of whether the storytelling task resulted in the generation of a fictional narrative (FICT), a personal narrative (PERS), or a narrative with both fictional and personal elements (MIXED).
3. Utterance completion was marked with end punctuation following the final end-of-utterance code. Incomplete or abandoned utterances were marked with ">".
4. Mazes: False starts, repetitions, hesitations, fillers and the like were offset with parentheses.

5. Unintelligible segments: Marked on the original transcripts with question marks, these were marked with one or more “X” consistent with SALT conventions.
6. Errors within the utterance: Utterances with any error as judged by a native English-speaking teacher were marked with an utterance level code “[EU],” which was placed at the end of the utterance before the end punctuation. Error types resulting in this code included grammatical, syntactical, morphemic, and word choice, errors. Pronunciation errors, which were indicated by phonetic spelling on the original transcript, were not counted as errors.
7. Subordination Index: an index of syntactic complexity, the subordination index counts the number of clauses per utterance.

Once coded, each file was converted to a plain text file and given “.slt” as a file extension so that it could be opened and analyzed with SALT software.

Narrative Scoring

Narrative macrostructure (organization) was evaluated by applying each of three different narrative scoring systems to the stories as described below. The scoring systems were applied to the corpus of stories one at a time, beginning with Story Grammar, the least complex system with a ten point holistic scale, then proceeding to Narrative Assessment Profile, which includes 8 categories each rated with a 3 point scale, and culminating with Narrative Scoring Scheme, which evaluates stories on 7 categories, each given a 5 point scale. All narrative analyses were conducted by the principal investigator with a subset of 20% of the narratives (n=17) also scored by a trained second

rater in order to determine interrater reliability, described in the section on reliability below.

Story Grammar Analysis

Using story grammar analysis (Hughes et al., 1997; Stein & Glenn, 1979), each story was examined for the presence of story grammar elements, which include setting, initiating event or problem, internal response, internal plan, attempt, consequence, resolution or reaction, and ending. Each narrative was then evaluated holistically on a ten-point scale to describe its overall organizational structure within a story grammar schema (e.g., reactive sequence, abbreviated episode, complex episode, etc.). The identification of elements and the assignment of a holistic score were guided by criteria provided by Hughes and colleagues. However, some modifications were made to the process of recording story grammar element information on a coding sheet.

Hughes and colleagues (1997) provide a coding sheet on which story grammar information is recorded by listing corresponding utterances to the right of a column in which each story grammar element is listed in a fixed temporal order. After attempting to use this format, which essentially required the scrambling of stories whenever they didn't conform to the fixed order, the PI and co-rater made the decision to create an alternate method. By using an Excel spreadsheet to record story grammar information, we were able to leave the original story intact with each utterance occupying a row in the leftmost column of the sheet while marking corresponding story grammar elements in a column to the right. Only those utterances deemed by the coder to fulfill a particular story grammar function in the narrative were ascribed an element. Some stories, namely those that had not minimally achieved the level of abbreviated episode (e.g., were designated as descriptive, action, or reactive sequences), were not coded for story grammar elements.

Importantly, story grammar elements are relative, not absolute, distinctions; they are meant to describe categorically the ways that specific kinds of information provided in a narrative function in relation to other specific types of information in order to construct a meaningful episode. Episodes necessarily entail goal-directed behavior of an identifiable main character or protagonist. In the absence of goal-directed behavior, story grammar elements lack meaning and are mostly undeterminable. The modified coding protocol also allowed each rater to reformulate utterances in order to better capture story grammar information and also allowed for the recording of comments and notes, which were helpful when ambiguities in the coding process surfaced.

Story complexity scores were assigned to each story by applying Stein's (1988) binary decision tree for determining story structure levels (Hughes et al., 1997) and by consulting Heilmann et al.'s (2010) ordinal adaptation of Stein's story levels. The latter was modified slightly by collapsing the first two complexity levels (isolated description and descriptive sequence) into one category (descriptive sequence). This decision was made because it was unclear to both the PI and co-rater what distinctions meaningfully separated the two categories. As we could find no adequate clarification in the literature, we chose to designate any narrative in which isolated characters, setting elements, and actions were described in no particular order as a level one story, or descriptive sequence. After descriptive sequence, story structure levels proceed with action sequences (actions chronologically ordered), and reactive sequences (a series of actions with some causal relations resulting in a chain of events, but with no evident goal-directed behavior). These first three levels of sequences are considered non-stories. Beginning at level four, goal-directed behavior is evident and stories begin to take the shape of episodes. Episodes may be abbreviated (level 4), incomplete (level 5), complete (level 6), multiple (level 7), complex (level 8), embedded (level 9) or interactive (level 10). Developmental

age has been empirically associated with each of these complexity levels. Children aged seven to eight are expected to generate stories at the levels of incomplete, complete, and multiple episodes (Hughes et al., 1997). For a detailed description of story grammar analysis criteria, see Appendix B1, the Story Grammar Analysis Decision Guide.

Narrative Assessment Profile

Based on their extensive research, Bliss, McCabe, and Miranda (1998) developed the Narrative Assessment Profile (NAP) specifically to evaluate discourse coherence related to both the macrostructure and microstructure of personal narratives. It focuses on six aspects of discourse coherence: topic maintenance, event sequencing, informativeness, referencing, conjunctive cohesion, and fluency. It has been adapted in numerous ways to meet specific research or clinical requirements. Each aspect represents a discrete dimension of the NAP. The instrument can be used both qualitatively and quantitatively. Whether data are recorded in a quantitative or qualitative fashion (McCabe & Bliss, 2003), the NAP essentially asks the evaluator to judge whether each dimension occurs with appropriate, variable, or inappropriate frequency in the context of the whole narrative. Appropriate occurrence in any given category is defined as that which is frequent enough so as not to reduce discourse coherence. Inappropriate frequency reduces discourse coherence and variable frequency occasionally reduces discourse coherence. To quantify the process, a scale is used and points are assigned to represent levels of appropriateness. Nevertheless, McCabe and Bliss caution that the story must be taken as a whole when assigning numeric values to levels of appropriate behaviors. Qualitatively, the evaluator is asked to describe the discourse patterns that are deemed to be variable or inappropriate. In this way, useful information is gathered that

will help the examiner identify areas of strength and need in a child's narrative development as well as inform foci for future interventions.

The NAP has been used to describe typical and disordered child narration across various cultural groups, including Spanish-speaking ELLs in America. The quantitative version of the instrument assigns greater weight to the category of informativeness by operationalizing it in three distinct ways typified by the following examples: 1) the police officer's needs to understand the incident the speaker is trying to relate (e.g., providing information that allows the listener to discern the gist of the story); 2) the teacher's goals for the speaker to give ample detail when describing an incident (e.g., providing elaboration of important details); and 3) the required "chef's ingredients" of action, orientation/description, and evaluation (the presence of all three is considered appropriate).

For ease of comparison with the other scoring systems, the quantitative version of the Narrative Assessment Protocol (McCabe and Bliss, 2003, pp. 175-176) was used. Each narrative aspect was given a rating of 0, 1, or 2, for 16 possible points. For each aspect, a score of 0 indicates either non-existent or mostly inadequate performance; a score of 1 indicates variable performance; and a score of 2 signifies complete and/or adequate performance. Comments were noted, as needed, to describe narrative discourse in each category, citing specific examples from each transcript as appropriate. For this analysis method, it is not necessary to analyze the text at the level of the utterance, rather stories are taken as a whole when determining the extent to which discourse patterns in each dimension contribute to or detract from the overall coherence of the narrative. Through the process of co-rating narratives to establish reliability, we modified slightly the NAP coding criteria provided by McCabe and Bliss to clarify decision rules that we

established upon discussing disagreements and ambiguities in the coding process. See Appendix B2 for the modified version of the NAP that we used.

Narrative Scoring Scheme

The Narrative Scoring Scheme (NSS), developed by the Language Analysis Lab at the University of Wisconsin-Madison, was designed to be a comprehensive and developmentally sensitive measure of narrative organization skills (Miller et al., 2006; Heilmann et al., 2010). It incorporates basic story grammar features but also narrative features characteristic of older, competent storytellers, such as a literate style of speaking through, for example, the use of metacognitive verbs to convey a character's thoughts and mental states. Additionally, the NSS evaluates the effective use of cohesive devices, including referential cohesion, conjunctive cohesion, and lexical cohesion. Special attention was paid to the scaling of the measure so that it would be capable of differentiating the narrative skills of younger and older children under a variety of narrative elicitation tasks.

The NSS is comprised of seven sections, each assessing a different aspect of narrative organization. Three of the sections, *introduction, conflict resolution,* and *conclusion*, are modeled after traditional story grammar analysis. The *mental states* and *character development* sections evaluate literate language skills, while the *referencing* and *cohesion* sections measure children's cohesion skills. Literate language is characterized by the use of metacognitive verbs, including verbs describing characters' thoughts and mental states, and metalinguistic verbs, which describe characters' speech. Cohesive language is characterized by the presence of and effective use of various cohesive devices. The NSS includes components evaluating literate and cohesive language because they are known to differentiate more mature and capable narrators from

novice narrators (Heilmann et al., 2010). Each narrative component on the NSS receives a scaled score of 0-5, resulting in a total score ranging from 0-35. The scaled score is determined by consulting a rubric (Heilmann et al., 2010), which describes minimal or immature performance (level 1), emerging (level 3), and proficient (level 5) performance for each narrative component. As with the other systems, through the process of establishing reliability with this instrument we made some modifications. Specifically, we found it helpful to add criteria to the rubric that described performance patterns in between the three levels of minimal, emerging, and proficient. See Appendix B3 for our modified rubric.

RELIABILITY

All narrative segmentation, scoring, and analysis were completed by the principal investigator. To be able to interpret the findings and speak to the trustworthiness of the coding process with some confidence, a second rater was enlisted to segment and score a randomly selected sample of 20% of the transcripts (n=17) using each system. In this section the process of training, scoring, resolving disagreements, and creating decision rules and other modifications to scoring protocols undertaken to establish adequate reliability with each of the methods is described. Final reliability coefficients calculated after all reliability coding was completed are reported here as well.

Selection of a Co-Rater

A co-rater was chosen to score a subsample of the study's narratives so that the reliability of the coding and scoring procedures could be measured and confirmed. The co-rater had an academic and linguistic background similar to that of the PI and the qualifications to accurately complete the task. Both the principal investigator and the co-rater have experience collecting and analyzing language samples of SS-ELLs and both

are native English speakers with proficiency in Spanish. Four distinct processes were subject to reliability testing on 20% of the sample's narratives prior to completion of data preparation and analysis by the PI. These processes were transcript segmentation, story grammar analysis, Narrative Assessment Profile, and Narrative Scoring Scheme, in that order. Each process was fully completed before the next process commenced.

Training Using Transcripts Generated by Select Non-Participant Sample

A sample of narrative transcripts that were collected from 2nd grade SS-ELLs under the same task conditions but who were not participants in the current study were used for training purposes. Ten narrative transcripts were purposefully selected to represent a range of narrative performance abilities similar to what would be encountered in the study's sample. During training for each of the coding processes, the PI provided the co-rater with detailed instructions and at least one sample narrative coded as an example. After discussing any questions, the co-rater and the PI each independently coded two sample narratives and results were compared. Disagreements and uncertainties were resolved by discussion. Resulting decision rules and modifications to help clarify existing instructions were recorded and made available to both coders for use during the next round of sample coding. This process continued until we demonstrated at least 80% agreement (calculated as number of agreed decisions divided by number of total decisions) on coding decisions and had dispelled any uncertainties related to our interpretations of criteria, at which point we proceeded to code the first few narratives in our reliability sample. The 80% agreement calculation was used to *estimate* agreement throughout the processes of training and modifying decision rules. Percent agreement was not used to determine actual reliability, which was calculated using Krippendorff's alpha (Hayes & Krippendorff, 2007), described in a subsequent section.

Selection of the Reliability Sample

The selection of the reliability sample occurred as follows. First, each narrative ID in the sample was listed in consecutive cells in a single column in an Excel spreadsheet. Each narrative ID was then ascribed a number in the column to the left such that the narratives could be identified with the numbers 1-83. A random number generator was then used to select 17 numbers within the range, 1-83. Those 17 narratives were used for each phase of reliability testing.

Process of Establishing Reliability

With each of the four processes for which reliability was sought and measured, it was necessary to make decisions about how to code phenomena that were not described a priori by the authors of the method under question. Thus the process of establishing reliability became one of maximizing the utility and reliability of each coding and scoring system by clarifying, where necessary, its language and criteria so that the system was easier to interpret and use with the kinds of narratives typical of our sample. All discussions of disagreements, clarifications and modifications to scoring criteria, and subsequent coding decisions were part of this process and were documented. Key issues and decisions are summarized as follows.

Transcript Segmentation

The first step in segmenting transcripts was to parse text into utterances. The original transcripts were often unpunctuated blocks of text so the first decision that had to be made was where to insert utterance breaks. Wherever intelligible simple sentences occurred, that decision was relatively easy. However, the presence of unintelligible segments (marked on the original transcripts with question marks) and strings of words that were disordered or that lacked any of the components of a complete clause required

that judgments be made based on guidelines that were generated through the process of consensus agreement. Other coding decisions that needed to be made included choosing end punctuation (marking whether the utterance was complete, abandoned, or interrupted), which words and chunks of text constituted mazes (marked with parentheses), how many clauses each utterance contained or whether that was indeterminable due to unintelligible segments, and whether there were any errors besides pronunciation errors within the utterance (marked with an end-of-utterance code, [EU]). Thus, while the narrative text presented many unambiguous instances where existing guidelines for coding utterances (Hughes et al., 2007; Miller et al, 2011) were applicable "out of the box", so to speak, there were also many situations requiring discussion and agreed rules for handling future situations of the same sort. The need for specific decision rules thus could not be anticipated before they were actually encountered during the process of coding utterances and so a set of rules evolved out of that process (see Appendix C).

Deciding what to code during transcript preparation was also a product of the process itself. While it was tempting to utilize all available SALT codes to generate maximum information about each transcript, decisions had to be made about which codes were necessary for the purposes of the current study and which could be left for future analyses. These decisions were made for the sake of expediency and to facilitate reliable coding. Ultimately, we were interested in reporting a few key statistics related to language productivity and narrative microstructure: the number of utterances, the mean length of utterances, the portion of grammatically acceptable utterances by native English speaker standards, and the complexity of utterances as measured by the presence of subordination. The process of establishing reliability of transcript segmentation resulted in a set of agreed narrative transcripts for the 17 narratives in the reliability sample and a

set of decision rules to guide the principal investigator in the preparation of the remaining 66 narratives.

Story Grammar Analysis

When transcript segmentation was complete, the agreed versions of the 17 transcripts were each pasted into an excel spreadsheet for coding of story grammar. The PI provided the co-rater literature describing story grammar and published guidelines for determining story grammar elements and for determining a story grammar organization level (Heilmann, 2010; Hughes et al., 1997). The PI also provided examples of 3 sample stories representing a range of performance patterns that were coded for story grammar. Upon reviewing the literature and discussing questions and disagreements regarding the first three sample narratives, the co-rater and the PI each independently coded 3 more sample narratives, after which disagreements were discussed and coding criteria modified. This process continued with sample narratives until at least 80% agreement on elements and holistic scores was reached and coding criteria were deemed adequately clear by both raters. We went on to code our set of reliability narratives independently, four narratives at a time. Upon scoring the first eight reliability narratives, we made some key revisions to the scoring guidelines based on the generation of decision rules in response to problems frequently encountered. We additionally agreed upon a set of procedures designed to facilitate consistency in our approaches to coding. As a result, we each recoded the first eight narratives with improved results exceeding our minimal standards for reliability (all results are reported in the Results section below). We then continued and coded the rest of the narratives with similarly acceptable results. See Appendix B1 for our revised Story Grammar Analysis Decision Guide. A summary of the issues leading to our revisions of the story grammar coding protocol follow.

The process of story grammar coding, particularly the coding of story grammar elements, required much discussion. There were some coding choices for which we determined that agreement could not be achieved because the decision to code a particular utterance as one particular story grammar element was contingent upon interpreting other utterances in a specific way. Where multiple interpretations were possible, multiple configurations of story grammar elements were also possible. All decisions ultimately rested on the coder's interpretation of whether or not the story had a protagonist and, more importantly, who that protagonist was. The choice of protagonist then affected what were considered to be the initiating events or problems and the subsequent attempts to address the problems. Indeed most of our initial disagreements during story grammar coding pertained to the identification of initiating events and attempts.

A few key modifications to coding criteria resulted in acceptably consonant decisions by raters. First, it was necessary to define a starting point for determining whether an analysis of story grammar elements was even appropriate for a given narrative. For story grammar analysis to be appropriate, goal-directed behavior must be evident (Hughes et al., 1997; Westby, 1992). The determination of whether goal-directed behavior is evident in the narrative thus became the first step in our modified decision tree. Our original guidelines, borrowed from Hughes and colleagues, suggested that story grammar analysis could be accomplished by first scanning the narrative for story grammar elements and then consulting a binary decision tree to determine whether the overall story organization meets criteria for each progressive story level beginning with level 1 (descriptive sequence). Our modifications required the coder to first read the story and decide whether or not goal-directed behavior was apparent. If it was not, then the story did not meet criteria to be considered an episode (level 4 and above) and could

only be considered a descriptive (level 1), action (level 2), or reactive (level 3) sequence. Story grammar analysis cannot be conducted with the first two levels and can only be minimally applied to a reactive sequence at level 3. This was an important distinction because, even at levels 1 and 2, some utterances may resemble story grammar elements and the temptation to code them as such may lead to erroneous conclusions. This is especially true wherever criteria suggest that the difference between two story levels depends upon the presence of one or more specific story grammar elements. For example, while it may be the case that certain elements (e.g., initiating event, attempt, and consequence) *must* be present for an episode to be considered "complete", it is not necessarily the case that the presence of all three guarantees that an episode is complete. The presence of these story grammar elements may thus be considered a necessary but not sufficient condition for ascribing to the story the level of complete episode. Consequently, if we start looking for story grammar elements without first determining whether or not a story actually exists, we risk inflating our estimation of a story's level as well as introducing opportunity for inter-rater disagreement.

The process of coding, however, demands that a decision be made about each utterance. The choice to not code an utterance because it doesn't fit available categories is still a choice, although it provides no information about how the utterance does function in the narrative. Because of this, as a second step in our revised decision-making protocol, we provided the option of categorizing non-narrative story elements, which according to Westby (1992) include actions, internal states, external states, and natural occurrences and are appropriate categories for descriptive and action sequences. Having the option to ascribe these categories to utterances was helpful in cases where the sequence resembled a true narrative yet lacked goal-directed behavior, for example, where natural occurrences served as problems, initiating reactions (not purposeful

behavior) in characters. Once a narrative was considered to possess goal-directed behavior, the next step was to identify the superordinate, or most important goal, and along with that, the protagonist. Stein (1982) suggests, and we concur, this is not always an easy task and is often a source of disagreement among raters. Story grammars are predicated on the introduction of a single protagonist around whose desires and goals episodes are organized (Mandler & Johnson, 1977; McCabe & Bliss, 2003). In the absence of a single identifiable protagonist or in the presence of multiple potential protagonists, there must be a way to select a main goal, thus the need for theories of importance to explain how a listener establishes superordinate goals. It is here, maintains Stein, that story grammars fall short. They generally lack an explanation for choosing one particular goal over another as being the most important. While there are some guidelines for establishing a main character (introduction at the very start of the story, greater number of utterances devoted to the character, character development through expressing mental states, etc.) and a superordinate goal (the ultimate goal of the main character), the selection is still a subjective process. With story grammar analysis, we are left with no real solution to the problem of multiple possible interpretations of a child's story with no way of confidently knowing which interpretation the child intended or what the child expects the listener to be able to infer. It is therefore difficult to judge the level of story organization that the child achieves. In order to increase the reliability of our interpretations, we decided to consider the first goal introduced as the superordinate one unless a goal presented later was clearly the one around which the narrative was organized.

Once a superordinate goal was determined in a goal-directed episode, we then proceeded to identify story grammar elements. The configuration of elements was then key to selecting an appropriate story level. For identifying story grammar elements we

referred to guidelines in Hughes et al (1997) and Westby (1992). For determining story levels, we relied upon Heilmann et al.'s (2010) ordinal adaptation of Stein's story structure levels as well as the levels described by Hughes and colleagues. We eliminated the first of Heilmann et al.'s levels (isolated description) because, as mentioned earlier, we felt it was not meaningfully distinct from the next level, descriptive sequence, which is the first story level described by Hughes and colleagues. We thus ascribed one of ten story levels to each story. We had little difficulty agreeing on story levels but found one to be somewhat problematic conceptually. The "multiple episode" ranks higher on an ordinal scale than does the complete episode, yet it is described by Heilmann and colleagues as a story that "contains more than one episode (either complete or incomplete)" (p. 622) and by Hughes et al as "a chain of reactive sequences or abbreviated episodes, or a combination of complete and incomplete episodes" (p. 121). Our criteria for determining whether or not a story was eligible for story grammar analysis required that goal-directed behavior be present, yet a series of reactive sequences, which has no goal-directed behavior, is not only eligible for story grammar analysis, it actually receives a higher score with the implication that it is either more sophisticated or well developed than a single episode that is complete. We considered changing the criteria to specify that a multiple episode would minimally consist of one complete episode plus one or more reactive sequences or additional episodes, complete or incomplete. We instead chose to abide by the existing criteria to see what would result when these criteria were applied to the study's sample of narratives. No further discussion of this issue is necessary in this section but will resume when results are discussed in chapters four and five.

Narrative Assessment Profile

Upon completion of story grammar analysis, the quantitative version of the Narrative Assessment Profile (NAP) (McCabe & Bliss, 2003) was provided the co-rater along with narrative samples scored by the PI. We repeated the process of discussing the criteria, independently scoring samples, discussing disagreements and revising the criteria to reflect agreed decision rules. All disagreements, discussions and changes were documented. What follows is a summary of key components of the process of becoming reliable using the NAP instrument.

The process of scoring the NAP differs from story grammar in that eight distinct categories are rated on a scale of 0-2 resulting in a cumulative score ranging from 0 to 16. The process required iterative readings of the narratives in order to evaluate each discrete category. Some of the categories (e.g., conjunctive cohesion, referencing) were fairly easy to interpret and we had little to no difficulty agreeing on scores. Other categories proved more troublesome. Among these were the three types of informativeness, fluency, and to a lesser degree, topic maintenance. Informativeness is weighed heavily in the NAP as it represents 3 out of the 8 categories. Informativeness is operationalized in 3 distinct ways, each of which is important to the comprehensibility of a narrative. Conceptually, the categories made sense to us. A good narrator provides information necessary to the police officer (the facts), information desired by the teacher (details, vivid description and elaboration), and the narrative "ingredients" essential to the chef (action, description, and evaluation).

While we generally agreed upon what constituted informativeness according to the police officer in any given narrative, we were less clear when evaluating the other types of informativeness. Namely, we needed to decide how to treat performance that was tangential. For example, if a child included rich descriptive detail about something

essentially off topic, would he or she be credited for appropriate performance for the narrative aspect of informativeness according to the teacher? The original criteria provided by McCabe and Bliss (2003) offered little help. They simply stated that appropriate performance (level 2) is "culturally apt elaboration;" variable performance (level 1) is "moderate elaboration;" and inappropriate performance (level 0) is "only 1-2 statements at best" (p. 175). In search of clarification, we consulted the general scoring guidelines applicable to all categories. The following were specified:

> Appropriate. A behavior is considered to be appropriate when the narrative behavior occurs frequently. Inappropriate behaviors are infrequent enough so as not to reduce discourse coherence.
>
> Inappropriate. A behavior is considered to be inappropriate when its frequency reduces discourse coherence.
>
> Variable. A behavior is considered to be variable when its frequency occasionally reduces discourse coherence but when the client shows some strengths on a particular dimension (p. 18).

Appropriate behavior (level 2) in any given narrative aspect is defined as behavior that occurs frequently enough (and we added, "and/or in the right proportion") so as not to reduce discourse coherence. Inappropriate behavior (level 0) is behavior whose frequency or infrequency reduces discourse coherence; and variable behavior (level 1) occurs when the behavior's frequency occasionally reduces coherence but the child shows some strength in the particular dimension as well. Viewing a discrete behavior in the context of its role in either supporting or diminishing discourse coherence was a helpful distinction and enabled us to agree that elaboration and other ingredients had to cohere with the topic. In fact, we encountered cases where children provided abundant detail about something in the picture (for example, the balloons in the circus picture or

what various characters were wearing in the baseball picture) and the amount of detail provided interfered substantially with the coherence of the story line. It made sense in most of these cases to rate the behavior as variable or even inappropriate, because although the child demonstrated the ability to elaborate, he or she chose to elaborate about some minor detail instead of providing information that enhanced the plot. Likewise, when evaluating children's informativeness according to the chef, the ingredients of action, description, and evaluation had to essentially support and not detract from the plot. The presence of description, evaluation and action statements that are off topic are often more compromising of discourse coherence than when such content is lacking.

A second area of ambiguity was fluency. The original criteria described appropriate performance as "fluent (in both languages)," variable performance as "a few dysfluencies," and inappropriate performance as "almost every utterance is dysfluent" (McCabe & Bliss, 2003, p. 176). The lack of specificity resulted in much disagreement, so we sought an empirical basis for establishing more objective criteria for rating fluency. After all, dysfluencies (mazes, hesitations, false starts, etc.) can be counted. In accordance with the general guidelines, we first established that an appropriate level of fluency should be one in which utterances are generally fluent and that any dysfluencies are so infrequent and so minor that they do not impede understanding. A variable level of fluency would indicate dysfluencies frequent enough that they occasionally interfere with discourse coherence. Performance would be considered inappropriate if most utterances are dysfluent to the point that comprehension breaks down. Damico, Oller, and Storey (1983) provided the criteria we sought. They examined the relationship between pragmatic difficulties and subsequent diagnosis of academically consequential language disorders in Spanish/English bilingual children ages 6 to 8. Pragmatic

difficulties, which included linguistic nonfluencies, revisions, delayed responses, nonspecific vocabulary, inappropriate responses, poor topic maintenance, and need for repetition (prompting) by the examiner, did indeed predict the language impairment status of bilingual children. Children exhibiting dysfluency in 30% or more of utterances in both languages were classified as language disordered. Based on these empirical data, we established that narratives with dysfluencies in over 30% of their utterances would be considered to evidence inappropriate performance. We defined variable performance as 20-30% of utterances with dysfluencies but still comprehensible, and appropriate performance as at least 80% of utterances exhibiting fluency.

Disagreements in topic maintenance were relatively few and were easily resolved through discussion. We felt the criteria were adequately clear and only added one clarifier, that to be appropriate *almost* all utterances needed to be on topic instead of all utterances. Given such a small scale (0-2), we wanted to be able to identify strong performance in the area of topic maintenance and felt that the occasional off topic utterance in an otherwise highly cohesive narrative shouldn't force the coder to designate performance as variable. It was too strict a standard, so we modified it.

For event sequencing, we needed to clarify that violations of sequence that were motivated (e.g., for emphasis) would not lower a narrative's score. On the other hand, a score of "0" would not necessarily mean "no" chronological ordering. Rather, it would depend to some extent on the length of the narrative and the overall portion that evidences appropriate event sequencing. We therefore added wording to reflect that inappropriate performance may be defined as either no chronological ordering of events *or* that most events are not in order. Furthermore, we clarified that when judging event sequencing, we needed to specifically consider the sequencing of on-topic, not off-topic, utterances. This requirement helped us establish a system for NAP coding such that we

started by looking for informativeness according to the police officer. We needed to first determine whether a story was present and then identify the topic, after which we evaluated the sequencing of information related to the topic. As it did with story grammar analysis, specifying an order to decision making increased our reliability when scoring the NAP.

Finally, we removed the wording in the original scoring protocol pertaining to culture. We did this for consistency sake. Since we relied on our judgments as native speakers of English to interpret scoring criteria and to guide our evaluations in all other coding activities, we wanted to do so similarly with the NAP, even though it does offer guidelines for considering culture when evaluating children's narratives. Since our sample is culturally homogeneous and because it is our intent to describe performance patterns, we felt it unnecessary to use the "culturally apt" qualifiers provided by the authors of the NAP. An attractive feature of the NAP is that it illustrates how cultural considerations can be accommodated within a scoring system. Because McCabe and Bliss (2003) reported variations along NAP dimensions on the performance of four ethnic groups, including Spanish-speaking Americans, their summary of findings will provide a valuable point of comparison when the results of this study are discussed.

Narrative Scoring Scheme

As with each of the other systems, training on the Narrative Scoring Scheme (NSS) began with providing the co-rater literature about the system, the rubric, and samples scored by the PI as an example. Questions were discussed and addressed and then we proceeded to independently score sample narratives. After the first round of scoring samples, we discussed disagreements and established agreement by consensus while adding clarifying language to the scoring rubric in order to make decision rules

adequately explicit. The NSS is similar to the NAP in that it examines several discrete aspects of narrative text and thus requires iterative readings in order to evaluate each aspect. The scale is broader (1-5) with level 1 defining minimal or immature performance, level 3 defining emerging performance, and level 5 defining mature or proficient performance. The authors of the rubric provide guidelines for identifying performance at levels 1, 3 and 5 for each characteristic. The levels represent a continuum or a progression of performance such that higher levels are additive of skills specified at the lower levels. For this reason, we found it best to work backwards when deciding on performance levels for each characteristic. Specifically, we began by questioning whether or not the narrative sample met criteria for a level 5 of a given characteristic. If not, we asked if it satisfied the requirements of a 3, and so on. Initially, we left levels 2 and 4 undefined, assuming that it would be simple enough to identify patterns of performance that fell between two defined levels. This turned out to be a faulty assumption and we eventually found it necessary to add to the rubric explicit criteria for levels 2 and 4 (see Appendix B3 for our final modified revision of the NSS rubric). We did this after scoring all 17 of our narratives with unacceptable levels of agreement. Once we amended the rubric we independently rescored all 17 narratives using the new and improved criteria and achieved acceptably high reliability (reported below). Disagreements were discussed and resolved by consensus and all decisions and issues encountered were documented. What follows is a summary of the key issues we encountered in becoming reliable in using this instrument and how each of those issues was addressed.

Some criteria provided on the original rubric were vague as stated and required specificity. For example, an emerging level of mental states was defined as “some use of evident mental state words to develop character(s)”. Minimal performance was defined

as “no use” and proficient performance was indicated when mental states were “expressed when necessary for plot development and advancement” and when “a variety of mental state words are used.” While “no use” is perfectly clear, the term “some” needed clarification. After some discussion, we agreed that the emerging category should reflect an attempt to elaborate mental state(s). We thus determined that a singular mention of one mental state could receive a rating of 2 but that to receive a 3, a narrative needed to have more than one mention of a mental state word. Later, we added a stipulation that a single *justified* mental state (which is a type of elaboration in which a reason for the mental state is provided, e.g., “he was scared because the lion was chasing him” or “he was scared so he started to run”) would also constitute emerging performance. That stipulation was added when we later attempted to define the difference between a 4 and a 5. A score of 5 indicated that mental states were “expressed when necessary for plot development and advancement”. We sought to operationalize this condition and agreed that the inclusion of mental states serves to advance and develop the plot when the mental states serve an explanatory function, explaining either a character’s behavior or internal reactions. Therefore, proficient performance does more than just state a character’s mental state(s); it links those mental states with actions and motivations that propel the plot. We amended level 5 to reflect this distinction and further specified that those connections must be clearly marked, not just implied. Thus a score of 5 would evidence a variety of mental state words elaborated with linking devices (because, therefore, so) while a 4 would come to be defined as a variety of mental state words used without explanations.

The category of character development was modified to provide clearer guidelines as well. Elaboration of characters is a hallmark of level 5, however we added that such elaboration needed to be sufficiently complete and not leave major questions or gaps in

the listener's comprehension. At level 3, both main and supporting characters are mentioned with no detail or description and main characters are not clearly distinguished from supporting characters. We ended up modifying level 3 upon clarifying level 1 and defining level 2. In all categories, revisions to one part of the continuum had ripple effects that necessitated slight and sometimes more substantive revisions to others points on the continuum. Level 3 thus moved from a mention of many characters with no ability to distinguish main from supporting characters (which now became a level 2) to mention of characters such that a hierarchy is evident and a main character is distinguishable from supporting characters. There is still very little detail provided, though, at a level 3. A level 4 was then defined as a clear attempt to develop a main character by elaborating, for example, that character's mental states and/or devoting a substantial number of utterances to that character.

Since character development begins with introduction, it was necessary to revise the introduction category as well. A proficient introduction would establish setting (time, place, context) while introducing at least one key character. These elements ought to be provided at appropriate places throughout the story. At the emerging level, attempts to establish setting are present but minimal. For example, a child may use a setting statement like "once upon a time" or provide a context for the story ("one day a boy was playing baseball"). We defined the emerging level as having 2 setting elements, one of which includes the introduction of an important character. A level 1 was the absence of any setting elements in which case a child launches into the story with no attempt to provide a context or to introduce characters. A level 2 thus became a singular attempt to address setting by providing either a setting (time/place) statement or introducing a main character. At level 4, setting elements and introduction of main character are present, but other characters are not adequately introduced or referenced.

The category of referencing evaluates the narrator's ability to enable the listener to maintain a clear understanding of who or what is being talked about. Poor referencing results in much confusion and places a considerable burden on the listener to repair gaps that interfere with comprehension. Referencing is a type of cohesion and, indeed, the two categories are interrelated. For evaluative purposes, however, they are distinct in that referencing specifically refers to the use of pronouns and antecedents while cohesion is concerned with event sequencing and transitions through the use of cohesive devices. The original criteria for an emerging level of referencing stipulated "inconsistent use of referents/antecedents." In an effort to more clearly define points on the continuum, we added that a level 3 includes "some appropriate and some inappropriate referencing but inappropriate referencing doesn't interfere with comprehension of the basic story; listener is not confused." Comprehensibility doesn't suffer although some references are inappropriate. This was necessary in order to establish a level 2, which also includes both appropriate and inappropriate referencing, however although some are appropriate, comprehensibility suffers. A level 1 was originally defined as the "excessive use of pronouns; no verbal clarifiers used; and the child is unaware that the listener is confused." We changed this to specify "no verbal clarifiers were used *when needed* or inappropriate use of articles (a/the) and other clarifiers (this, that, these, those, etc.)." A proficient level of referencing was indicated by no ambiguity pertaining to characters and, we added, the appropriate use of clarifiers (e.g., "*the first* boy," or "*the other* boy") to enhance comprehensibility. We added the appropriate use of clarifiers to level 5 in order to distinguish it from a level 4, which would simply indicate no ambiguity pertaining to characters.

We went on to define the levels of cohesion. At the low end, we defined minimal cohesion as "a series of disconnected utterances," basically a descriptive sequence. At a

level 2, there is an attempt to connect utterances with cohesive devices (e.g., “and”). Level 3 or emerging performance with cohesion was indicated by the presence of a logical order of events. At this level, however, excessive detail or emphasis placed on minor events may lead the listener astray, or equal emphasis on all events (through, for example, restricting conjunctions to only “and” or “then”) prohibits the distinction of important events from non-important ones. At a level 4, we expected to see some appropriate use of subordinating conjunctions and at a level 5, events follow a logical order, critical events are emphasized, and smooth transitions are provided between events.

The original criteria for conflict resolution and conclusion were modified similarly in order to provide better definition of performance at each level. The two categories are interrelated in that a story can never be fully concluded unless important conflicts are resolved. But conclusion goes beyond conflict resolution using devices that signal not only the resolution of an event but also the end of narration. In some cases a narrator will signal the end of narration (e.g., “that’s all,” or “the end”) without providing a resolution for an important event. In other cases, there is no conflict or resolution yet a concluding statement marking the end of narration is provided. Therefore the conclusion category has its own set of criteria distinct from conflict resolution. Through discussion of disagreements, we settled on some decision rules to help clarify existing criteria so that they could be interpreted and applied to the narratives more reliably. Namely, to achieve a proficient rating on conflict resolution, not only must all conflicts and resolutions critical to advancing the plot be clearly stated, the resolutions provided must be adequate and must not leave the listener hanging. At the emerging level, original criteria stated that “not all conflicts and resolutions critical to advancing the plot are present.” We added, “there is at least one discernible conflict and resolution but they may be under

developed." At the minimal level, we separated original criteria between level 1 and level 2. We established a level 1 as having no discernible conflict whereas a level 2 would be assigned to a narrative that mentioned a conflict but not a resolution or a resolution but not a conflict. Level 4 is more advanced than a level 3 in that all *critical* conflicts and resolutions are present but it is less than level 5 in that they may be under developed (e.g., not entirely explicit) and require some logical inference on the part of the listener.

While the NSS was the most difficult system on which to become reliable, it also prompted some interesting observations and reflections, documented during reliability scoring and conversations over disagreements. After the first round of four reliability narratives were scored, I made the following comments in my notes:

> We both agree that these are tricky to score. However, the process does bring up some interesting information and questions about each child's narrative abilities. It forces us to look deeper than our surface impressions of the story. Most of the kids are actually scoring higher than I would have given them if I were to rate the entire story holistically and not by its component parts.
>
> Cultural considerations emerge. For example, in narrative 3 (11061_52708), the resolution by my standards is unclear; however, there is enough said that an actual resolution can be inferred. If the child had worded the final utterance just a little differently it would have changed the meaning enough to be clear and not leave a question in my mind. Is this stylistic? In other words, does his choice of words reflect a tendency toward a more collectivist, high-context communication style in which it is assumed that the listener shares the speaker's general knowledge/orientation and that inferences such as this one are to be expected?

Measuring Inter-Rater Reliability

Reproducibility of a coding instrument is a determination of its reliability that involves evaluating whether different observers of a set of phenomena (e.g., a set of oral narrative transcripts of second grade Spanish-speaking ELLs), given common instructions, yield similar results within a tolerable margin of error (Hayes & Krippendorff, 2007). Krippendorff's Alpha has been recommended as the most appropriate reliability statistic for content analyses (Gwet, 2011; Hayes & Krippendorff, 2007). Krippendorff's alpha satisfies each of the conditions for a good index of reliability by calculating disagreements instead of correcting percent-agreements, which is a limitation of many of the other commonly used reliability statistics (Hayes & Krippendorff, 2007). Furthermore, it is capable of measuring comparable agreements for nominal, ordinal, interval and ratio data. According to Krippendorff, alpha values greater than or equal to .80 are adequately reliable and values between .67 and .80 are acceptable for exploratory research and for drawing tentative conclusions. For the purposes of this study, an agreement level of .80 was sought while a minimal agreement level of .67 was considered acceptable.

Using a macro (Hayes & Krippendorff, 2007) for Krippendorff's alpha (KALPHA) written for the Statistical Package for the Social Sciences (SPSS) (IBM Corp, 2011), KALPHA coefficients were calculated for two types of agreement: agreement between the two raters; and agreement between the PI and consensus decisions. Results for each type of agreement are reported below for each coding activity: 1) transcription segmentation and coding; 2) story grammar analysis; 3) Narrative Assessment Profile; and 4) Narrative Scoring Scheme.

Transcript Segmentation

Krippendorff's alpha was calculated for the number of coded utterances, number of mazes, subordination index, and the number of utterances with errors. Alpha coefficients between PI and agreed decision exceeded .96 for all categories (see Table 3.2).

Table 3.2

Krippendorff's Alpha Coefficients for Reliability of Narrative Segmentation

	Between PI and Co-rater	Between PI and Agreed Decision
Number of coded utterances	0.9598	0.9878
Number of mazes	0.8934	0.9678
Subordination Index (SI) count	0.9788	0.9934
Utterances with Errors (UE) count	0.7599	0.9656

Note. The reliability sample consisted of 17 narratives randomly selected from the study sample of 83. The principal investigator (PI) and one co-rater segmented each of the 17 narratives. Disagreements were resolved by discussion resulting in a third "agreed decision."

Story Grammar Analysis

Story grammar elements coded dichotomously as either present or not present in a given narrative. Percent agreement was calculated for story grammar elements, resulting in 100% agreement between PI and co-rater as well as PI and agreed decision for each of the elements. The story grammar holistic scores assigned by each rater were compared and Krippendorff's alpha was calculated to measure agreement, which was .96 between PI and co-rater as well as agreed decision (see Table 3.3).

Table 3.3

Percent Agreement and Krippendorff's Alpha Coefficients for Reliability of Story Grammar Analysis

	Between PI and Co-rater	Between PI and Agreed Decision
Story Grammar Elements*		
Setting (Set)	100%	100%
Initiating Event/Problem (IEP)	100%	100%
Internal Response (IR)	100%	100%
Internal Plan (IP)	100%	100%
Attempt (A)	100%	100%
Consequence (Con)	100%	100%
Resolution (Res)	100%	100%
Ending (E)	100%	100%
Story Grammar Holistic Score	0.956	0.9567

Note. Story Grammar elements were dichotomously coded for their presence or absence in each narrative, thus simple agreement expressed as a percentage is provided, as this measure is appropriate when used with categorical data generated by two coders. Story Grammar holistic score is ordinal on a scale of 1-10, thus Krippendorff's alpha was used to express agreement.

Narrative Assessment Profile

Each rater's scores for each of the eight categories of the NAP were compared, as were their total composite scores. Agreement was highest for the category of fluency, for which there was no disagreement. Each reliability transcript received a score of 0 from each rater in the fluency category; therefore alpha could not be calculated. It was next highest for conjunctive cohesion at .97. Agreement for the rest of the NAP categories ranged from .67 (informativeness teacher) to .96 for the composite scores. Informativeness teacher is perhaps the most subjective of all the categories and thus the most difficult for which to achieve agreement. Alpha levels are reported in Table 3.4.

Table 3.4

Krippendorff's Alpha Coefficients for Reliability of Narrative Assessment Profile

	Between PI and Co-rater	Between PI and Agreed Decision
NAP Categories		
Event Sequencing	0.878	0.878
Informativeness: Police Officer	0.9179	0.9179
Informativeness: Teacher	0.6708	0.6708
Informativeness: Chef	0.8734	0.8734
Referencing	0.7195	0.778
Conjunctive Cohesion	0.9689	0.9689
Fluency	NA	NA
NAP Total Score	0.9563	0.9576

Note. Each NAP category is rated on a scale of 0-2, resulting in an NAP total score with a scale of 0-16. Krippendorff's alpha was used to measure agreement between raters on each coding decision. There was no variation present in the ratings on fluency (all values were agreed to be 0), thus alpha is undeterminable.

Narrative Scoring Scheme

The NSS scores for each of its 7 categories were compared as well as the composite score. Alpha levels ranged between .87 and 1 between PI and agreed decision and between .86 and .99 between PI and co-rater. These highly acceptable alpha levels were achieved with the revised NSS, which resulted from our modifications during the reliability process.

Table 3.5

Krippendorff's Alpha Coefficients for Reliability of Narrative Scoring Scheme

	Between PI and Co-rater	Between PI and Agreed Decision
NSS Categories		
Introduction	0.9357	0.9357
Character Development	0.9609	0.9609
Mental States	0.9949	0.9949
Referencing	0.8745	0.8873
Conflict Resolution	0.9752	0.9752
Cohesion	0.8613	0.8729
Conclusion	0.9203	1
NSS Total Score	0.9369	0.9564

Note. Each NSS category is rated on a scale of 0-5, resulting in a NSS total score with a scale of 0-35. Krippendorff's alpha was used to measure agreement between raters on each coding decision.

DATA ANALYSIS

Data Preparation

Data were organized into spreadsheets and then imported into SPSS where variables were redefined as necessary to accurately represent data types (nominal, ordinal, or interval) prior to analyses.

Descriptive Statistics

Descriptive statistics were generated on all data through SPSS reporting functions. These include the distribution and frequency of scores, means and standard deviations for each measure. These statistics describe and compare participants' narrative organization skills according to each of the scoring systems.

Answering Research Questions

The research questions guiding this investigation are repeated here, after which the methods for answering each question are described.

The following research questions guide the study:

1. What are the characteristics of second grade Spanish-speaking ELLs' stories, orally narrated in English, using the following methods of analyses?
    a. Story grammar analysis
    b. Narrative Assessment Profile
    c. Narrative Scoring Scheme
2. How does each scoring system characterize the sample in terms of expected narrative performance according to its own criteria?
3. What are the distinguishing features of narratives whose scores are consistent (e.g., high, average, and low) across measures?

4. What features must a narrative scoring system have in order to provide teachers with quality information that will help them design instruction and interventions for SS-ELLs?

After all data were prepared and compiled into a dataset in SPSS, a series of reports and analyses were generated to address the research questions. Beginning with microstructural characteristics, means were generated for each measure with picture prompt as a factor. Means were compared using analyses of variance (ANOVA) to test for significance of differences at the $p < .05$ alpha level. With each scoring system, composite score means and distributions were first described followed by means and the frequency of scores for each of the measure's subcomponents. Once each scoring system's results were described, the sample was stratified along the distribution of NSS scores because that measure produced the most normal distribution of the three. Average performers were defined as those whose NSS scores fell within a range of one standard deviation above or below the mean for the sample. In the same way, below average and above average groups were established. Stratification by one standard deviation (rather than by two) was chosen because of the relatively small sample size. Because the objective of stratification was to identify and describe characteristics of narrative performance at each of the levels, it was necessary to have a sufficient number of narratives representing low, average, and high performers to be able to detect patterns of performance at each level.

Upon stratification by NSS scores, individual cases were then categorized accordingly and case summaries were reported with NSS classification as a factor to determine how average, below, and above average narratives fared with the two other measures. Cut points were established for story grammar and for NAP scores to establish average, below, and above average groups according to each of their respective criteria

for expected performance. Each narrative in the sample was thereby given a categorical rating for each scoring system. From these categorical ratings a subset of narratives whose ratings were consistent (average, low, or high) across scoring systems were identified. This subset was used to identify and describe the stable features of performers. Features of scoring systems were described as well and discussed in terms of their ability to recognize and appropriately evaluate these stable features.

SUMMARY OF METHOD

The purpose of the study is fourfold: (a) to describe the characteristics of the English stories of Spanish-speaking ELLs; (b) to compare how each scoring system characterizes the sample according to its own criteria; (c) to identify the stable features of narratives rated consistently across scoring systems; and (d) to identify criteria for a high quality narrative scoring system for evaluating the English oral narratives of Spanish-speaking ELLs. To fulfill its purpose, a reliable method for analyzing and scoring the sample of narratives with each scoring system was sought. The PI and a co-rater applied a rigorous process of co-rating non-sample narratives first to test and revise each scoring system's rubrics and criteria. Disagreements were resolved through discussion and consensus while decision rules were generated and refined. Modified rubrics resulted and were applied to a subsample of 20% (n=17) of the sample's narratives to determine the reliability of the modified systems. Reliability was measured by Krippendorff's alpha. Wherever results were unacceptable, further consensus decisions were made and modifications to the rubrics ensued, after which the reliability sample was rescored by each rater. Only after acceptable reliability levels were achieved did the PI continue to score and analyze the sample's remaining narratives. Reliability levels in most cases exceeded .80, which is considered sufficiently reliable. Analysis methods included the

comparison of means of composite and subcomponent scores and involved the stratification of the sample to identify stable performance features.

CHAPTER FOUR

Results

Good narrative organization skills are associated with literacy and with academic achievement in general. The inability to form a coherent narrative even when one's language skills (e.g., vocabulary, fluency, grammar) are adequate for the task is characteristic of monolingual individuals with learning disabilities (LD). In bilingual individuals with LD, this inability will manifest itself in both the native and the second language. While it is ideal to compare narrative samples in both languages to assess a bilingual individual's narrative skills, unfortunately this is not an option for many of the SS-ELLs attending U.S. public schools. Because most ELLs are taught in English-only settings, methods are needed for evaluating SS-ELLs' performance in English. Evaluation of narrative organization is useful to evaluate for SS-ELLs because metalinguistic abilities are discernible by quality measures of narrative organization, despite the surface language errors and dysfluencies common throughout the process of acquiring a second language. Furthermore, a quality narrative measure will provide useful information to teachers of SS-ELLs, not only about their students' metalinguistic narrative organization abilities, but about other language strengths as well. For example, the resources they employ to make an extended narrative discourse coherent, cohesive, and meaningful when they have limited English vocabulary and grammar.

The prevailing methods for evaluating children's oral narrative performance have mostly been developed and tested on monolingual English-speaking populations. Cultural and linguistic differences may affect how Spanish-speaking ELLs perform on such measures, creating the potential for systematically skewed results and erroneous interpretations when narrative organization scoring systems are used with SS-ELLs. The

purpose of this investigation is twofold: (a) to describe the characteristics of SS-ELLs' oral English narratives by identifying performance patterns, as measured by three different scoring systems, in a sample of fictional narratives generated in response to one of two picture prompts; and (b) to identify criteria for a high quality scoring system that is sensitive to both the strengths and the specific areas requiring improvement in SS-ELLs' oral English narratives. Performance patterns are described by identifying the stable features of the children's narratives when they are analyzed and evaluated by each of the three measures. With these stable features as a standard for what SS-ELLs' oral English narratives are truly like, the particular aspects of each scoring system that are well suited and those that are mismatched to reflect those patterns are identified. The identification of criteria for a high quality system is accomplished by comparing the ways that each of the systems sort SS-ELLs into low, average, and high performers and also provide information about the relative strengths and weaknesses of their performances. A high quality system ought to accomplish three things: (a) sort children systematically, reliably and meaningfully into categories of average, low, and high performers; (b) identify those who are potentially in need of intervention; and (c), provide specific information about each child's strengths and weaknesses across the most important aspects of oral narration, specifically those contributing to discourse coherence. Consequently, the scoring systems used in this study will be scrutinized in two ways: (a) how they categorize performance across the sample; and (b) the qualities of information they provide about each narrative's strengths and weaknesses.

To achieve the study's purposes, the following research questions guide the investigation:

1. What are the characteristics of second grade Spanish-speaking ELLs' stories, orally narrated in English, using the following methods?

a. Story grammar analysis
b. Narrative Assessment Profile
c. Narrative Scoring Scheme

2. How does each scoring system characterize the sample in terms of expected narrative performance according to its own criteria?
3. What are the distinguishing features of narratives with consistent scores (e.g., high, average, and low) across measures?
4. What features must a narrative scoring system have to provide teachers with quality information that will help them design instruction and interventions for SS-ELLs?

Research question one is addressed by describing the characteristics of the sample's narratives. General characteristics are described followed by specific results of microstructural and macrostructural analyses. To answer research question two, aggregate characteristics are described in terms of each system's distribution of total and subcomponent scores. For research question three, the narratives were stratified by their performance on the Narrative Scoring Scheme, whose distribution was the most normal of any of the measures, and divided into three groups: average performance (those narratives whose scores fell within one standard deviation (SD) above or below the mean), below average performance (scores less than 1 SD below the mean), and above average performance (scores greater than 1 SD above the mean). For the other two scoring systems, cut levels were established for average, below, and above average performance based on criteria specific to each rubric. Those categorical ratings were then compared to identify the subset of narratives rated similarly by each scoring system. Stable, or consistent features of each category were then identified and described. Discrepant ratings were examined for patterns that reveal features of scoring systems that

are mismatched with the characteristics of the sample or of the other systems and which may render certain scoring systems less well suited to the analysis of SS-ELLs' oral English narratives. Of special interest were those features of systems that consistently rated SS-ELLs' narrative performance as low when they were identified as average or above average on the NSS. Research question four is answered by summarizing the findings related to those criteria that were particularly well suited to providing quality information useful to instructional planning with SS-ELLs.

THE CHARACTERISTICS OF SPANISH-SPEAKING ENGLISH LEARNERS' NARRATIVES

General Findings

There were significant ($p < .05$) differences between the amounts of language produced in response to the two different picture prompts (pictures are provided in Appendix A). The circus picture resulted in significantly longer stories with more vocabulary used as measured by number of total utterances (NU), number of total words (NTW), and number of different words (NDW). However, examining the results of macrostructural analyses, there were no differences in the overall organization of the stories when comparing the means associated with each picture prompt for any of the three scoring systems. This finding suggests that SS-ELLs are able to construct a story in English even with limited vocabulary. It further suggests that narrative organization scoring systems are able to detect quality in an SS-ELL's English narrative regardless of its length.

Microstructural Characteristics of SS-ELLs' Narratives

During the process of segmenting narrative transcripts, raters judged whether the story adhered to the picture prompt, in which case the story was coded as "prompt driven," whether it partially adhered to the prompt ("partially prompt driven") or did not

adhere to the prompt ("not prompt driven") or was "not determinable". Additionally, the child's interpretation of the task as judged by the type of story delivered (fictional, personal, or mixed) was recorded. Table 4.1 illustrates that 94% of the stories were prompt driven, 5% were partially prompt driven, and 1% was not determinable.

Table 4.1

Narrative Picture Prompts and Prompt Adherence

	Prompt			
	Baseball	Circus	ND	Total
ND	0	0	1	1
PD	38	40	0	78
PPD	4	0	0	4
Total	42	40	1	83

Note. PD = prompt driven; PPD = partially prompt driven; ND = not determinable.

The subjects interpreted the storytelling task appropriately, with 81 of the 83 stories appearing to be fictional. One child told a story in which fictional and personal content were mixed and one story's interpretation was undeterminable. See table 4.2.

Table 4.2

Narrative Task Interpretation

	Prompt			
	Baseball	Circus	ND	Total
ND	0	0	1	1
FICT	42	39	0	81
MIXED	0	1	0	1
Total	42	40	1	83

Note. ND = not determinable; FICT = story was fictional; MIXED = story included both fictional and personal (anecdotal) elements.

Narratives varied greatly in length. Table 4.3 reports language productivity for narratives elicited in response to each picture prompt. The range of language produced is substantial, from two utterances to 64 utterances and from 8 words in the analysis set (e.g., the set of completed and intelligible utterances and the set of words excluding mazes) up to 553 words. To get a better estimate of means, I excluded the one most extreme outlier in terms of numbers of words and utterances (553 words and 64 utterances as compared with the next highest values of 296 and 37, respectively). Additionally, I excluded four anomalous cases identified by SPSS as the 5% cases with the highest anomaly index values (5% is the default set by SPSS). These exclusions resulted in a set of 78 narratives, including 41 baseball and 37 circus narratives. New means were calculated for number of utterances in the analysis set and number of total words (see Table 4.4).

Table 4.3

Mean, Median, Minimum and Maximum Values of Language Productivity Variables

	Baseball (n=42)	Circus (n=40)
Number of Utterances (Analysis Set)		
Mean (Std. Error)	8.76 (.811)	13.4 (1.75)
Median	7	9
Minimum	2	2
Maximum	22	64
Mean Length of Utterance (MLU)		
Mean (Std. Error)	6.98 (.223)	7.27 (.272)
Median	6.8	6.79
Minimum	3.67	4
Maximum	10.57	12
Number of Different Words (NDW)		
Mean (Std. Error)	33.79 (2.48)	44.53 (3.96)
Median	31.5	35
Minimum	8	8
Maximum	65	119
Total Words in Analysis Set (NTW)		
Mean (Std. Error)	62.26 (6.26)	100.88 (14.97)
Median	55.5	68
Minimum	11	8
Maximum	159	553
Utterances with Errors in Analysis Set (UE)		
Mean (Std. Error)	3.48 (.389)	6.63 (.954)
Median	3	6
Minimum	0	1
Maximum	11	39
Subordination Index (SI)		
Mean (Std. Error)	1.17 (.035)	1.19 (.032)
Median	1.145	1.145
Minimum	.88	.86
Maximum	2.17	1.75

Note. Number of Utterances (analysis set) = number of main clauses and their arguments (analysis set includes only those utterances that are complete and intelligible; abandoned, incomplete, and unintelligible utterances are not included in the analysis set); Mean Length of Utterance = average length of utterance in words; Number of Different Words = an index of vocabulary representing the number of unique words used in the narrative; Number of Total Words = calculation of total words used in the utterances included in the analysis set; Utterances with Errors = utterances coded as having an error in grammaticality or word choice unlikely to be made by a native English speaker; Subordination Index = an index of sentence complexity calculating number of main clauses plus subordinate clauses divided by number of main clauses (utterances).

Table 4.4

Mean, Median, Minimum and Maximum Values of NTW and NU, Outliers and Anomalous Cases Excluded

	Baseball (n=41)	Circus (n=37)
Number of Utterances (Analysis Set)		
Mean (Std. Error)	8.95 (.813)	12.11 (1.21)
Median	7	9
Minimum	3	5
Maximum	22	37
Total Words in Analysis Set (NTW)		
Mean (Std. Error)	63.51 (6.28)	88.73 (9.61)
Median	56	63
Minimum	11	29
Maximum	159	296

Tables 4.3 and 4.4 both document differences in the amount of language produced in response to each picture. Analyses of variance (ANOVA) were conducted to determine whether differences were significant at the $p < .05$ level of confidence. Table 4.5 illustrates that in response to the circus picture prompt, children generated significantly more language in terms of number of utterances, number of words, number of different words, number of total words after accounting for mazes, number of mazes, and number of utterances with errors. This was true even when outliers and anomalous cases were excluded (see Table 4.6). Variables that did not differ by picture prompt include mean length of utterance, length of mazes in words, ratio of utterances with errors to total utterances in the analysis set, and the subordination index. Therefore differences

appear to be present only with respect to the amount of language produced, not the quality of language in terms of its complexity and grammaticality. These findings suggest that the children had more content knowledge of and vocabulary with which to narrate about the circus scene. Nevertheless, the stories they constructed related to the baseball scene, of which they appeared to have less knowledge and/or words to draw from, were of similar quality, both organizationally and syntactically.

Table 4.5

Comparisons of Means (All Cases) of Select Productivity Measures by Picture Prompt

	Means				
	Total (n=82)	Baseball (n=42)	Circus (n=40)	F	Sig
No. Utterances (Anal. Set)	11.04	8.79	13.4	5.912	**0.017**
Total Words	121.24	91.24	152.75	8.731	**0.004**
MLU	7.12	7.27	6.97	0.71	0.402
NDW	39.02	33.79	44.53	5.407	**0.023**
NTW	81.1	62.26	100.88	5.86	**0.018**
No. of Maze Words	18.85	14.64	23.28	6.27	**0.014**
Words per Maze	1.85	1.77	1.94	1.88	0.174
Utter. with Errors (UE)	5.01	3.48	6.63	9.67	**0.003**
UE Ratio	0.5	0.45	0.55	2.564	0.113
Subordination Index	1.18	1.17	1.19	0.127	0.722

Note. No. of Utterances (Anal. Set) = total number of utterances that were complete and intelligible and thus subject to analysis; Total Words includes all words while number of total words (NTW) refers to those words included in the analysis set; Mean length of utterance (MLU) = NTW divided by number of utterances in the analysis set; number of maze words = total mazes while words per maze = average length of maze strings, which are offset in parentheses; utterances with errors include utterances with word order errors, verb tense errors, pronoun errors, etc. – any type of error except pronunciation errors; UE ratio = ratio of utterances with errors to total utterances (utterances with unintelligible segments are excluded); subordination index = main + dependent clauses/main clauses.

Table 4.6

Comparison of Means (Outliers and Anomalous Cases Excluded) of NTW and NU by Picture Prompt

	Means with Outliers Excluded				
	Total (n=78)	Baseball (n=41)	Circus (n=37)	F	Sig
No. Utterances (Anal. Set)	10.45	8.95	12.11	4.862	**.030**
NTW	75.47	63.51	88.73	5.008	**.028**

Note. Number of utterances and number of total words are the two most direct measures of the quantity of language produced. These two variables were selected to illustrate that outliers were not responsible for the significant difference in means between lengths of stories for each picture.

Macrostructure Characteristics of Narratives

In contrast to microstructural analyses, there were no differences in means on any of the macrostructural measures for stories generated about the two different pictures. This suggests that despite having more language with which to tell a story about the circus picture, children told stories of comparable organizational quality and coherence. The interpretation of mean scores is discussed in subsequent sections, however means are reported here for the sake of comparison. Mean scores for story grammar (scale of 1-10) were 3.48 for baseball and 3.98 for the circus picture. Median scores for both pictures were 3. Means for the NAP (scale of 0-16) were 6.33 and 6.22 respectively, with medians of 6 and 6.5. NSS (scale of 0-35) means were 19.02 and 19.03 with a median of 18 for baseball and a median of 20 for the circus picture. Results are reported in Table 4.7 and Figures 4.1 and 4.2.

Table 4.7

Story Grammar (SG), Narrative Assessment Profile (NAP), and Narrative Scoring Scheme (NSS) Mean Scores by Picture Prompt

	Baseball (n=42)	Circus (n=40)
Story Grammar Raw Score		
Mean (Std. Error)	3.48 (.33)	3.98 (.35)
Median	3	3
Minimum	1	1
Maximum	9	10
Narrative Assessment Profile Raw Score		
Mean (Std. Error)	6.33 (.63)	6.22 (.51)
Median	6	6.5
Minimum	0	1
Maximum	14	14
Narrative Scoring Scheme Raw Score		
Mean (Std. Error)	19.02 (.93)	19.03 (.86)
Median	18	20
Minimum	9	7
Maximum	31	31

Note. Story grammar scale 1-10; NAP scale 0-16; NSS scale 0-35.

Figures 4.1 and 4.2 illustrate mean raw scores and mean percentage scores for stories told in response to each picture prompt for each narrative organization measure.

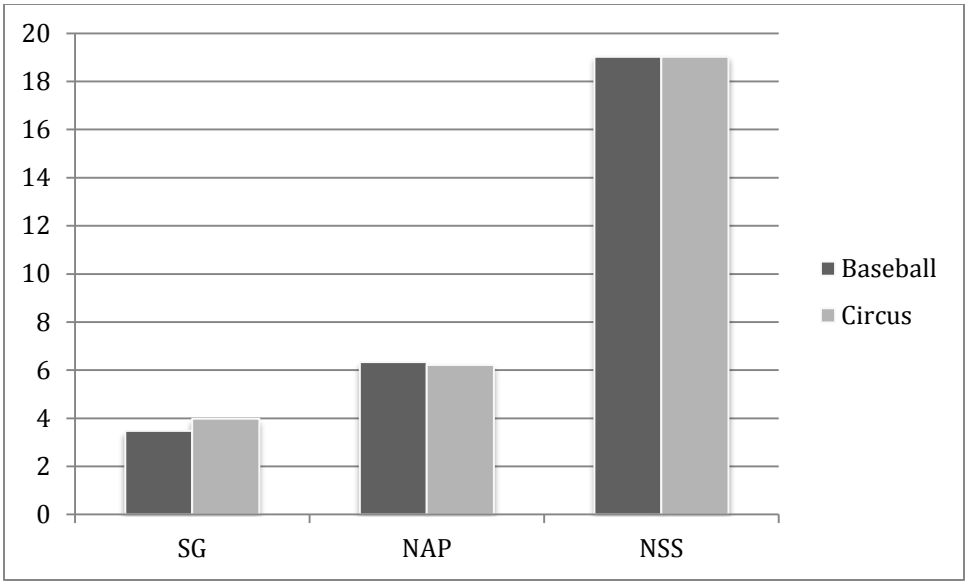


Figure 4.1. Mean raw scores achieved on narrative organization measures. SG = story grammar (scale of 1-10); NAP = Narrative Assessment Profile (scale of 0-16); NSS = Narrative Scoring Scheme (scale of 0-35).

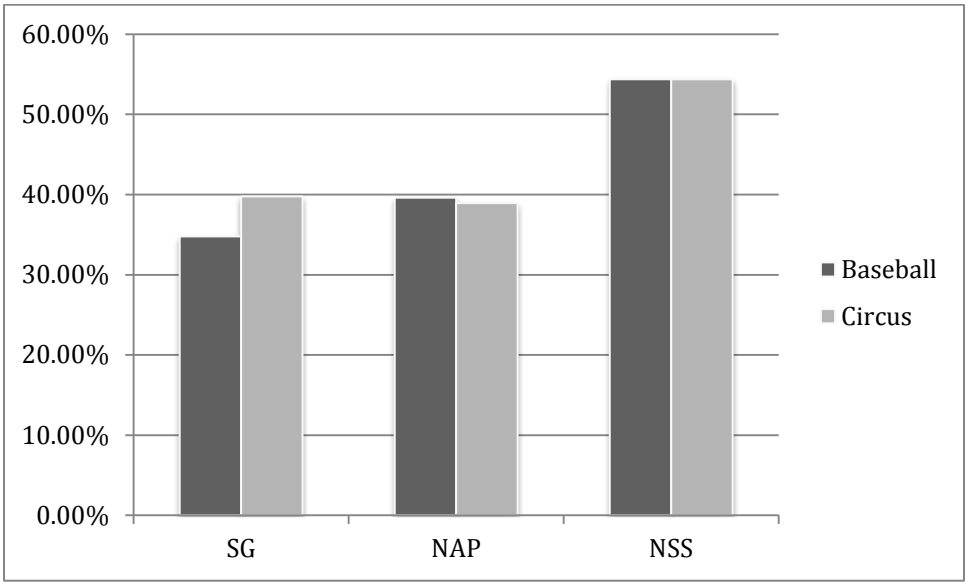


Figure 4.2. Mean percentage scores achieved on narrative organization measures. SG = story grammar; NAP = Narrative Assessment Profile; NSS = Narrative Scoring Scheme. Percentages were calculated by dividing the mean score by the total possible score for each scoring system.

Analyses of variance (ANOVA) confirmed that there were no differences in total scores on narrative organization measures for stories told in response to each picture (see Table 4.8). Because of this, further analyses of narrative organization measures do not consider picture prompt.

Table 4.8

Comparison of Mean Scores across Picture Prompts on Narrative Organization Measures

	Means			ANOVA	
	Total (n=82)	Baseball (n=42)	Circus (n=40)	F	Sig
Story Grammar	3.72	3.48	3.98	1.104	.297
NAP	6.28	6.33	6.23	.018	.895
NSS	19.02	19.02	19.03	.000	.999

Each narrative scoring system evaluated narrative performance somewhat differently, resulting in sets of characteristics, some of which can be compared between systems and some of which are unique to each system. Characteristics are first described according to how they surfaced through each method of scoring and analyzing the narratives, and both commonalities and differences are discussed in the next sections.

Characteristics of SS-ELLs' Oral English Narratives as Measured by Story Grammar Analysis

Holistic story grammar analyses rely on a binary decision scheme to determine which ordinal level of a scale the story achieves. The story grammar schema used in this study had a 10-point scale, the first 3 levels of which indicated sequences that lacked goal-directed behavior and thus were classified as non-narratives and were exempt from story grammar coding. Sequences were either descriptive (level 1), action, implying

some chronologically ordered events (level 2), or reactive, implying some causally connected events (level 3). Beginning at level 4, episodes emerged in abbreviated fashion and increased in completeness and complexity as scale scores increased. A notable exception to the order of increasing completeness and complexity is a level 7 narrative, which may be defined as a series of reactive sequences or some combination of episodes and reactive sequences.

According to story grammar analyses, the vast majority of the sample's stories are categorized as reactive sequences (level 3). In fact, 28 of the 83 stories (33.7%) were coded as reactive sequences. If one adds to that number those narratives coded as multiple episodes (level 7), which were predominantly series of reactive sequences, the portion of reactive sequences increases to 45.7%. 27.8% of the stories were categorized as action or descriptive sequences, leaving only 26.5% of the sample's stories as constituting true "narratives", according to story grammar. See Table 4.9 for the frequency and Figure 4.3 for the story grammar holistic score distribution.

Table 4.9

Frequency of Story Grammar Holistic Scores

Story Grammar			
Score	Frequency	Percent	Cumulative Percent
1	12	14.5	14.5
2	11	13.3	27.7
3	28	33.7	61.4
4	12	14.5	75.9
5	3	3.6	79.5
6	2	2.4	81.9
7	10	12	94.0
8	3	3.6	97.6
9	1	1.2	98.8
10	1	1.2	100.0
Total	83	100.0	

Note. 1 = descriptive sequence; 2 = action sequence; 3 = reactive sequence; 4 = abbreviated episode; 5 = incomplete episode; 6 = complete episode; 7 = multiple episode; 8 = complex episode; 9 = embedded episode; and 10 = interactive episode.

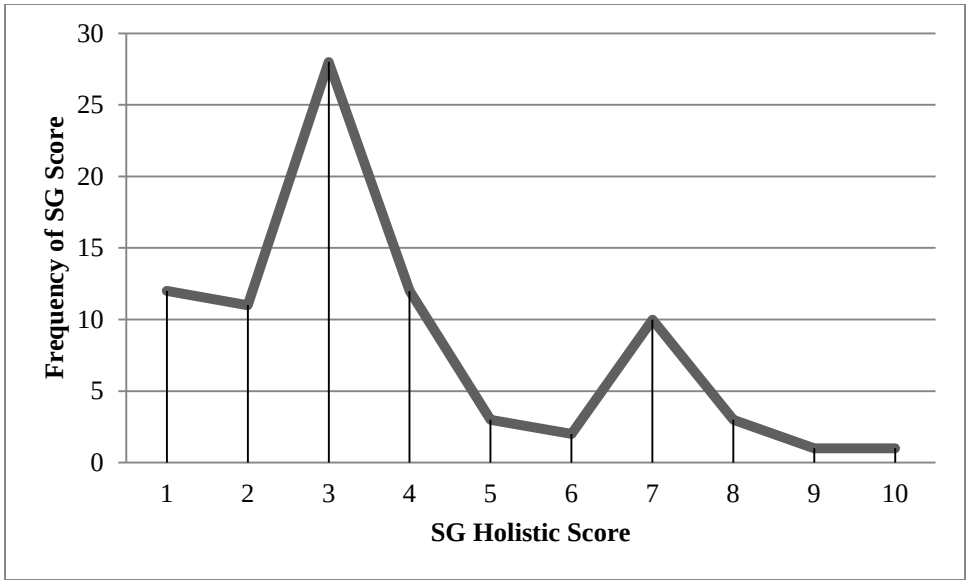


Figure 4.3. Story grammar holistic score distribution. 1 = descriptive sequence; 2 = action sequence; 3 = reactive sequence; 4 = abbreviated episode; 5 = incomplete episode; 6 = complete episode; 7 = multiple episode; 8 = complex episode; 9 = embedded episode; and 10 = interactive episode.

Reactive sequences, in which some story grammar elements may be identifiable, are characterized by a series of causally linked events without the presence of goal-directed planning or behavior. They are qualitatively different than descriptive or action sequences, because those are generally isolated utterances appearing to be randomly ordered, with the exception that an action sequence will evidence a plausible temporal order. Reactive sequences are *near* narratives and may in many ways resemble a story. From the framework of story grammar schema, the missing element in a reactive sequence is an attempt. So while there may be characters that experience problems and consequences, there are no explicit attempts on the part of a main character to solve those problems.

This is an example of a reactive sequence:

"One day a boy (uh) XX (a a) a ball house XX in…

(da da) da person calls to police.

And then the police catch the boy.

(and) and the police go to the jail (for the) for the boy."

Even though the first utterance is incomplete and unintelligible, the rest of the utterances are sufficiently comprehensible to be able to identify a causally connected series of events. It almost appears to be a story because there is an attempt at a resolution (the police go to the jail for the boy, or takes the boy to jail), however there is no goal directed behavior on the part of the presumed protagonist, the boy. This is more than an action sequence, however, because each event sets off the next event in a kind of chain reaction. Because of its brevity, it is likely that most teachers and other listeners would have no hesitation rating it as a pre-narrative.

In contrast, some narratives rated as reactive sequences were much better developed, or potentially so, yet still did not meet criteria to be considered an episode. The following example typified this type of story.

"There was this kid on a circus XX very happy.

(He was about to get) he was about to get a lot of balloons.

(And he wa) and something was checking him out.

It was a lion, a very angry lion.

He didn't know.

He thought he was a little bit XX.

Then it turns out his XX angry because he escaped (from from the) from the (ca) cage he was in.

And the lion (fall) fall back.

He dropped his popcorn XX his popcorn.

Everybody was tripped out XX except for the little boy.

He was scared of the XX a lot of the XX.

I hope this boy is XX.

I hope he is gonna be XX.

The lion (go) runs real fast.

And he has (big) big (um) hands.

((Oh gosh)).”

This story, which didn’t turn out to be a story at all by story grammar criteria, has some qualities that set it well apart from the typical narratives in this sample. Specifically, the introduction has a certain sophistication that is atypical, using language that engages the listener. Nevertheless, to meet the requirements to be considered an abbreviated episode or higher, goal directed behavior would need to be present and there is no identifiable goal. Further, the cohesion evident at the beginning comes apart as the child slips into commentary about the scene toward the end. The fact that both of these narratives qualify for the same score of 3 according to story grammar analysis, however, suggests that story grammar may be ill-suited to capture certain nuances of narrative performance.

Besides assigning a holistic score, story grammar analysis involves quantifying the number of story grammar elements present in each narrative that is rated 4 or higher. Of the narratives (n=32) that met this criterion, the elements of consequences, settings, and internal responses (IR) (e.g., mental states) were included with the greatest frequency (see Table 4.10).

Table 4.10

Average Number of Story Grammar Elements in Narratives at Level 4 and Higher (n=32)

	N	Min	Max	Mean	St. Dev.
Setting	32	1	7	2.438	1.664
IEP	32	1	5	1.75	1.016
IR	20	1	7	1.90	1.553
IP	0	--	--	--	--
Attempt	26	1	5	1.769	1.210
Consequence	32	1	6	2.938	1.390
Resolution	22	1	1	1	.000
Ending	8	1	2	1.125	.3536

Note. IEP = initiating event/problem; IR = internal response; IP = internal plan.

Attempts and initiating events/problems (IEPs) were the next most frequently included elements, followed by endings and resolutions. No narratives included internal plans. Another way of describing these patterns is that all of the narratives rated as abbreviated episodes and higher possessed the elements of setting, IEP, and consequence. Goal directed behavior is signaled explicitly by the presence of internal plans, of which there were none, and implicitly by the presence of attempts and/or internal responses. Not all of the 32 narratives coded for story grammar had internal responses and attempts, however each had one or the other.

Consequences can be related either to the attempts of characters (e.g., 'he swung the bat' [attempt], 'but he missed' [consequence]), or directly to the IEP. In the case of the latter, a problem is posed and consequences directly follow with no attempts or plans for action by a specific character. An example of this type of consequence would be "the

lion gets out of the cage [IEP] and then the boy throws his popcorn [consequence] and he runs [consequence]." However, if the child adds, "So the lion won't get him," the proposition, "and he runs" becomes an attempt: "and he runs so the lion won't get him." The inclusion or exclusion of the statement, "So the lion won't get him," which certainly could be inferred if not stated, may result in altogether different holistic ratings according to story grammar analysis.

Resolutions are a special kind of consequence. They are the ultimate consequence of the episode and they effectively put an end to the action. A good resolution means a listener will know that the story's end has been reached. It will not leave a listener hanging, wondering what happened. Examples of resolutions include, "And the police caught the boy and took him to jail," or "the man got the lion and put him back in the cage." Twenty-two episodes provided a resolution and 8 provided either one or two statements that additionally signaled the end of the narrative by concluding, for example, "and he never played baseball again," or "and the boy never went back to the circus again."

Characteristics of SS-ELLs' Oral English Narratives as Measured by the Narrative Assessment Profile (NAP)

The NAP examines eight aspects of oral narratives, including topic maintenance, event sequencing, informativeness (operationalized as the police officer's needs for the facts of the experience, the teacher's need for elaboration, and the chef's needs for the three "ingredients" of description, action, and evaluation), referencing, conjunctive cohesion, and fluency. It rates each of the eight narrative aspects on a scale of 0 to 2 with 0 representing inappropriate performance, 1 indicating variable performance, and 2 indicating appropriate performance. The composite score is the sum of each of the subcomponent ratings, ranging from 0 to 16. The sample mean was 6.22 out of 16

possible points, indicating that narrative performance according to the NAP was, on average, rated as “inappropriate” to “variable”. Degrees of appropriateness are not determined by developmentally typical behavior, rather by the criterion of discourse coherence. Behavior that is appropriate (receiving a score of 2) occurs frequently, contributing to discourse coherence with any inappropriate behaviors occurring infrequently enough so as not to reduce discourse coherence. Behaviors that are considered to be inappropriate (receiving a score of 0) are those whose overall frequency reduces discourse coherence. Variable behaviors (score of 1) are those whose frequency occasionally reduces discourse coherence, but where strengths are evident in that particular narrative dimension as well. Table 4.11 and Figure 4.4 illustrate the frequency and distribution of NAP composite scores, respectively.

Table 4.11

Frequency of NAP Composite Scores (n=83)

NAP Composite Score	Frequency	Percent	Cumulative Percent
0	1	1.2	1.2
1	8	9.6	10.8
2	6	7.2	18.1
3	9	10.8	28.9
4	4	4.8	33.7
5	8	9.6	43.4
6	10	12.0	55.4
7	12	14.5	69.9
8	5	6.0	75.9
9	2	2.4	78.3
10	7	8.4	86.7
11	2	2.4	89.2
12	2	2.4	91.6
13	3	3.6	95.2
14	4	4.8	100.0
Total	83	100.0	

Note. NAP composite scores may range from 0 – 16. No story achieved a score of 15 or 16.

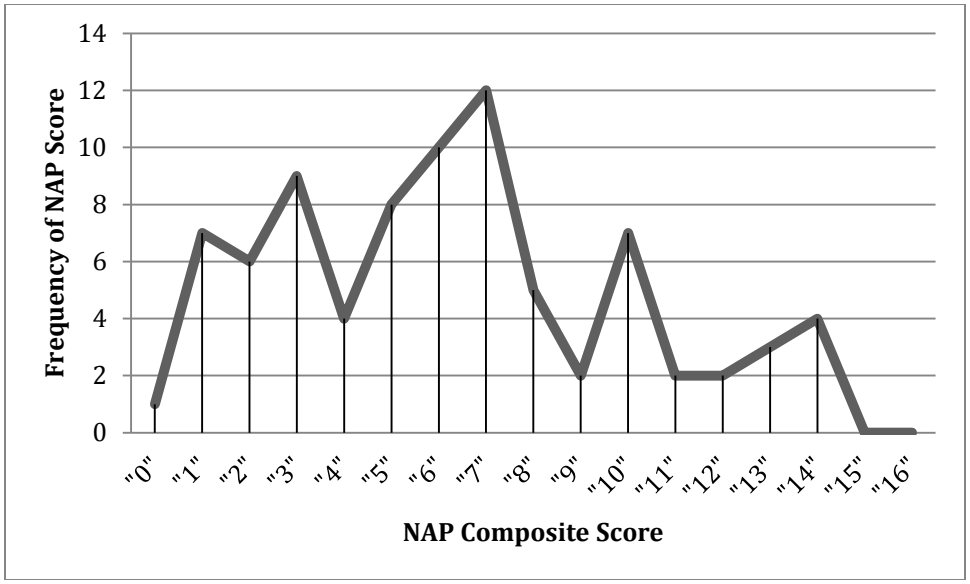


Figure 4.4. Narrative Assessment Profile total score distribution.

Of each of the NAP's eight subcomponents, conjunctive cohesion received the highest mean rating followed by topic maintenance, indicating mostly variable to appropriate performance in these categories. The categories of event sequencing, informativeness according to the police officer, referencing, and informativeness according to the teacher each reflected inappropriate to variable performance. Informativeness according to the chef and fluency received the lowest ratings, reflecting mostly inappropriate performance (see Table 4.12).

Table 4.12

Average Scores of NAP Subcomponents

	N	Min	Max	Mean	St. Dev.
Topic	83	0	2	1.024	.781
Event Sequ.	83	0	2	.916	.752
Info PO	83	0	2	.855	.701
Info TCH	83	0	2	.614	.696
Info CHEF	83	0	2	.373	.657
Referencing	83	0	2	.759	.820
Conj. Coh.	83	0	2	1.639	.508
Fluency	83	0	1	.036	.188

Note. Topic = topic maintenance; Event Sequ.= event sequencing; Info PO = informativeness according to the police officer; Info TCH = informativeness according to the teacher; Info CHEF = informativeness according to the chef; Conj. Coh. = conjunctive cohesion.

In Figure 4.5, subcomponents are ordered from left to right along the horizontal axis from those with the fewest number of inappropriate ratings to the greatest. This facilitates recognition of the sample's relative strengths and weaknesses. The frequency of scores occurring in the sample reveals that, according to the NAP, the children were fairly evenly distributed in terms of their performance on topic maintenance and event sequencing with notably fewer demonstrating appropriate levels of informativeness. Conjunctive cohesion was an area of strength; most subjects' performance was appropriate, as demonstrated by the appropriate use of a variety of conjunctions (and, then, but, so, etc.). Fluency was an area of weakness for the entire sample. Dysfluency was defined as 30% or more of utterances possessing mazes. Considering the sample of utterances in aggregate, nearly half (n=412) of all analyzed utterances (n=908) possessed

mazes. The actual rate of dysfluency is even higher if abandoned and partially unintelligible utterances are included in the calculation.

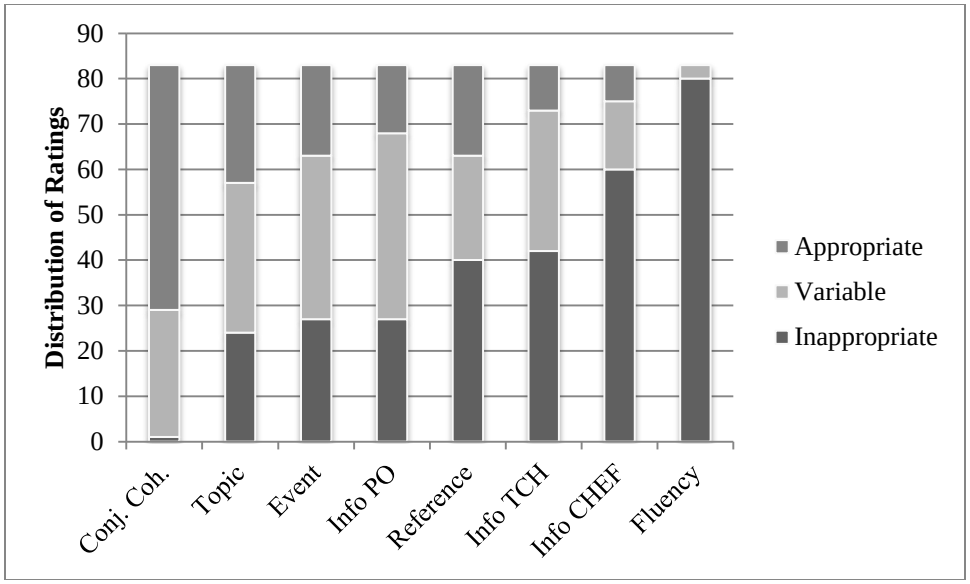


Figure 4.5. Frequency of ratings for each of the NAP subcomponents. Conj. coh = conjunctive cohesion; topic = topic maintenance; event = event sequencing; info PO = informativeness according to the police officer (gist information); reference = referential cohesion; info TCH = informativeness according to the teacher (embellishment and elaboration of details); info CHEF = informativeness according to the chef (inclusion of description, action, and evaluation in proper measure); fluency is determined by the proportion of mazes to total words (inappropriateness is defined as 30% or more of the total number of words being mazes; variable is defined as between 20%-29%; appropriate is less than 20%).

Informativeness is weighted heavily in the NAP, constituting 6 out of the potential 16 points on the scale. Its three categories are meant to evaluate three aspects of discourse coherence, specifically the types of information needed to make sense of the narrative (McCabe & Bliss, 2003). These include the kinds of information a police officer would request (e.g., the important facts of the experience being relayed), the kind of information a teacher would request in order to make the narrative engaging to

listeners (e.g. embellishment of details), and the basic narrative "chef's ingredients" of description, action, and evaluation in the right proportions. All three kinds of information contribute to discourse coherence. Factual information enables the listener to understand the gist of the story; embellishment of important points aids the listener so that they do not need to infer essential details; and the presence of each of the narrative ingredients ensures that attributes of people and objects are described, events are conveyed, and the significance of events is communicated. Systematic omission of any of these aspects reduces narrative coherence. Of the three aspects of informativeness, the SS-ELLs in this sample were relatively more competent at providing information that helps the listener understand the gist of the story (essential facts). Embellishment of details (informativeness according to the teacher) was mostly inappropriate to variable in the sample as was the inclusion of the three ingredients of description, action, and evaluation (informativeness according to the chef).

The NAP quantitative version, which was used to score the narratives in this study, provides general guidelines for scoring each of its subcomponents. Some of the narrative subcomponents (e.g., conjunctive cohesion) have very clear, objective guidelines for what constitutes inappropriate, variable, and appropriate behavior. For some of the other subcomponents, however, inappropriate, variable, and appropriate ratings each may be inclusive of a variety of behaviors. This is the case with the three aspects of informativeness. More information is needed to describe the ways in which SS-ELLs' English oral narratives are sufficiently or insufficiently informative. For example, an inappropriate rating on informativeness according to the teacher (Info TCH) can result from either the narrator providing little to no embellishment of important details or from providing too much embellishment of unimportant details.

To capture this information, I added columns in the coding spreadsheet (and subsequently variables in the SPSS data set) and returned to the 17 reliability narratives to apply a more nuanced analysis of the informativeness of those narratives. This additional analysis was undertaken as an attempt to explore the potential of the instrument to generate more specific information by adding coding categories. The findings of this effort ought therefore to be interpreted with caution, as there was no reliability procedure established for the additional analysis and it treated only a subsample of the narratives.

For informativeness according to the police officer (Info PO), I added the three variables of "context" (the who and where of the story), "problem" (what is it about), and "outcome" (what ended up happening). For each of those variables, I dichotomously coded them as providing insufficient information (the presence of important gaps) or providing sufficient information. For Info TCH, I added one variable to code for the specific reason a given narrative was rated as inappropriate or variable: Either the narrative exhibited a) not enough detail, b) some detail but still some important gaps, or c) too much detail that was off topic. Finally, for informativeness according to the chef (Info CHEF), the subcomponent score of 0, 1, or 2, is based on the number of ingredients (description, action, and evaluation) that are included. But this does not provide any information on *which* ingredients are included and which are excluded. Therefore, I added 3 variables: one for description, one for action, and one for evaluation. For each ingredient, I determined whether it was a) not present, b) present but with important gaps, c) present in too great a portion (excessive emphasis given to that one ingredient), or d) if that ingredient was included in appropriate measure. See Table 4.13 for a summary of the coding criteria that were added to provide better detail about children's performance.

Table 4.13

Additional Coding Variables and Criteria for NAP Informativeness

Additional Coding Criteria
Info PO Original 2 = All specific information necessary to understand experience is provided or implied; credit should be given for easily inferred information (A) 1 = Most specific information provided but omissions of a few important points (for example, beginning, middle, or end), - leaving the listener with some questions as to what happened (V) 0 = Not enough information (or too much information) to make sense of what happened (I) Info PO Additional For each of the following, determine whether information provided was insufficient or sufficient: • *Context* (who, where, etc. – information regarding characters, setting, etc.) • *Problem* (What is story about? What is the problem or issue?) • *Outcome* (What happened in the end?)
Info TCH Original 2 = Provides elaboration and embellishment of most important points of story including at least 2 of 3 ingredients (evaluation includes inferencing) (A) 1 = Provides some elaboration of some important points including at least 1 of the 3 ingredients (V) 0 = Provides little to no elaboration (1-2 statements at best) OR provides too much elaboration of extraneous details, detracting from storyline (I) Info TCH Additional If narrative was rated as 0 or 1, specify why: • No or very little elaboration • Some elaboration but still some gaps where elaboration is needed to increase coherence • Too much elaboration on unimportant details
Info CHEF Original 2 = All three ingredients must be present and appropriate in proportion; they provide sufficient information, leaving no gaps and/or creating no "noise" that impedes understanding (A) 1 = Two ingredients are present without important gaps (V) 0 = One or no ingredients are present without important gaps (I) Ingredients are: description (orientation), action, and evaluation Info CHEF Additional For each of the ingredients: • Description • Action • Evaluation Which of the following apply? • No information • Some information but gaps • Too much information • Appropriate information

Based on the more nuanced analysis of the subsample of the 17 reliability narratives, some patterns emerged. In providing factual information to convey the gist of the story (Info PO), children more consistently provided information regarding the problem statement and to a lesser extent, context, and were most likely to leave gaps regarding the outcomes of events. The listener was often left wondering what happened in the end. With respect to elaboration (Info TCH), all three types of inappropriate behaviors (providing no elaboration, providing some elaboration with gaps, and providing too much elaboration of unimportant details) were present in fairly equal portions. No particular patterns stood out. Informativeness according to the chef (Info CHEF) provides additional information about the kinds of ingredients (and thus the types of elaboration) present. According to Info CHEF analyses, the ingredient of description (elaboration of the attributes of characters and objects) was consistently sparse. By contrast, evaluation was abundantly present, and in some cases excessive, in these children's narratives. Action was consistently provided, however most stories had important gaps in action information, as evident, for example, in the failure to provide some final action that resolves the event, signaling its conclusion.

The omission or underdevelopment of specific types of information has repercussions on other narrative aspects and, ultimately, the overall coherence of the narrative. When gaps are present in information about the action in a story, the coherence related to event sequencing suffers. When descriptive details are sparse, referencing suffers. Indeed, even in stories that were otherwise well elaborated and well constructed, it was often difficult to keep track of who was doing or experiencing what because character descriptions were virtually nonexistent requiring that the listener sort out who was "the boy" mentioned in one utterance as opposed to "the boy" referred to in another utterance. Consider the following text from one of the narratives (mazes are

offset in parentheses and unintelligible segments are indicated by XX; references to the two main characters are underscored):

> "One time it was (a) a mom (and) and (a little bb bo) a little boy (who) who go and play basketball. One day the little boy was crossing (and) and somebody was running because he was his friend. And (he said) he said the police grab his hand and say, 'You're too little (to) to cross (the the) the street. You need to cross when (we, uh) all the people was seeing him, the little boy.' One day when the boy going (to) to play basketball, the police see him and grab his hand and he crosses (his his the street) the street, or the 'nother street, and (he said) the police say to the little boy, 'Never cross (the) the street with your friends. You're so little, you need to cross XX the street with your mom or dad or a big sister.' (Um) then his friend run away (and and) with his father. Then (he he told, he he) he told the everyone what happened to the little boy. Then the little boy never (cross crossed the street) crossed the street with his friend. (He needed) he every single day cross the street with his mom or dad or sister or brother."

This example is a fairly well developed story, which received high ratings across each of the systems for its overall completeness, organization, and complexity. Nevertheless, the listener or reader is burdened with distinguishing between the two boys. There are some notable strengths in referencing. The narrator does provide some appropriate antecedent and pronoun use as well as some clarifiers in referring to "the little boy" and "his friend". The narrator even self corrects and repairs an ambiguous reference when he or she clarifies, "…and he crosses….the street… and (he said) *the police say* to the little boy…" Nevertheless, referencing cohesion would be greatly improved if the narrator were to provide some additional details about the characters at the start of the story and use those to distinguish the characters throughout. For example,

if the narrator were to specify, “the little boy with the blue cap” and “his friend, the one with the red cap”, the additional details would provide a means to keep the characters distinct as the action becomes complex, as it does in this story.

Similar patterns are evident in many of the stories in the sample. Due to vague referencing and underdeveloped descriptions of characters, the listener is often left wondering, is the narrator referring to the same boy or a different boy, and how many boys are there actually in the story? Likewise, pronouns were used excessively, often without clear antecedents. Poor to variable referencing diminished overall coherence in a majority of narratives, placing a considerable burden on the listener to make inferences in order to repair gaps in the information provided. Its ability to identify not only the performance deficits (e.g., reference cohesion) but also specific strategies to improve those deficits (e.g., provide more detailed descriptions of characters) is a particular strength of the NAP with the revisions described in Table 4.13.

Characteristics of SS-ELLs’ Oral English Narratives as Measured by the Narrative Scoring Scheme (NSS)

The NSS rates narratives across 7 domains using a scale of 0 to 5 where: 0 is used only when a narrative may not be scored due to interruptions or errors in the examination process; 1 indicates minimal or immature performance; 3 indicates emerging performance; and 5 indicates proficient or mature performance in a particular domain. Composite scores are the sum of subcomponent scores and may range from 0 to 35. NSS composite scores of this sample ranged from 7 to 31 (out of 35 possible points) with a mean of 18.904 (standard deviation = 5.8) (see Figure 4.6 for distribution of scores).

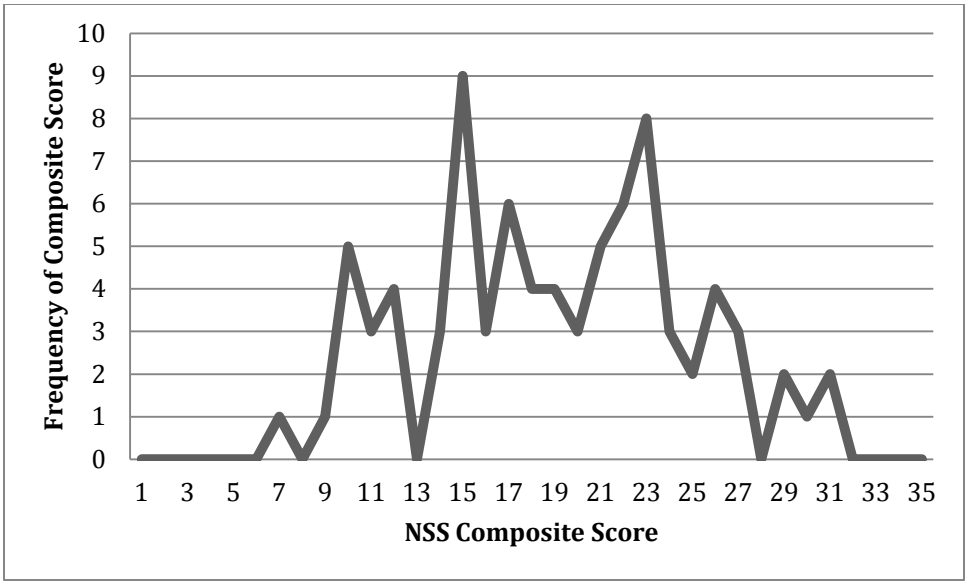


Figure 4.6. Narrative Scoring Scheme total score distribution.

Mean scores on NSS subcomponents clustered just below the level of emerging competency (level 3) for all seven subcomponents except character development, whose mean was 3.253. Character development and cohesion were thus the relative strengths of the sample according to the NSS, however performance levels were fairly equivalent across categories, ranging between a mean of 2.458 (mental states) to 3.253 (character development) (see Table 4.14).

Table 4.14

Average Scores of NSS Subcomponents

	N	Min	Max	Mean	St. Dev.
Intro	83	1	5	2.614	1.069
Char. Dev.	83	1	5	3.253	.9858
Mental	83	1	5	2.458	1.328
Reference	83	1	5	2.506	.9800
Conflict Res.	83	1	5	2.675	1.149
Cohesion	83	1	5	2.916	.8440
Conclusion	83	1	5	2.482	1.282

Note. Intro = introduction; Char. Dev. = character development; Mental = mental states; Conflict Res. = conflict resolution.

Mean NSS subcomponent scores indicate emerging competence across categories. Frequencies of scores (see Figure 4.7) illustrate how many narratives were characterized by performance patterns described on the NSS Scoring Modified Rubric (Appendix B3).

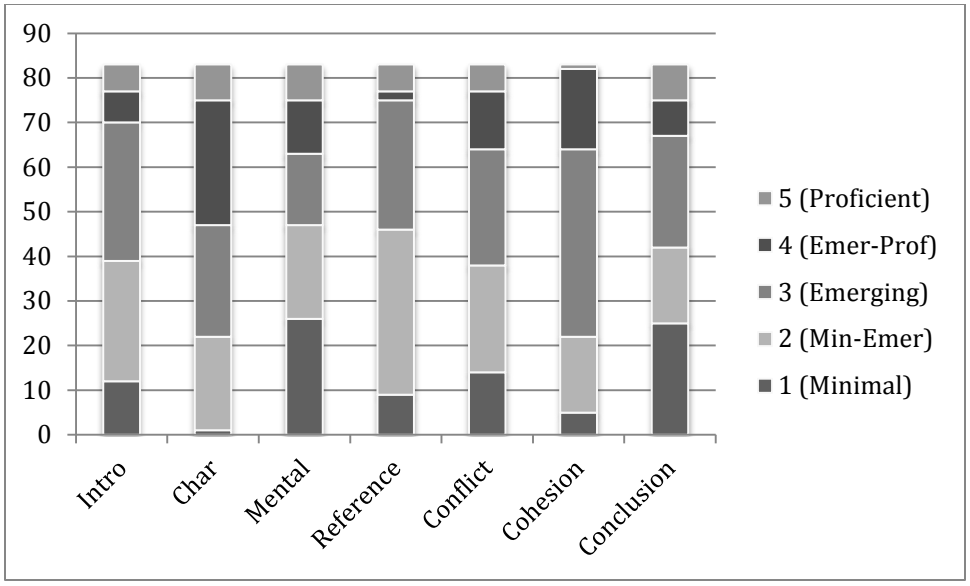


Figure 4.7. Frequency of ratings for NSS subcomponents across the sample (n=83). Intro = introduction; Char = character development; Mental = mental states; Conflict = conflict resolution.

To make sense of these patterns, however, one must consider the criteria for achieving various levels according to the rubric. The NSS was designed to evaluate narrative organization (macrostructural) skills associated with traditional story grammar analysis as well as language skills (microstructural) characteristic of a literate way of speaking (Heilmann, Miller, Nockerts & Dunaway, 2010). The NSS categories of introduction, conflict resolution, and conclusion are meant to capture those macrostructural features of narratives associated with story grammar's episodic structure, roughly corresponding with the story grammar elements of setting, IEP and resolution, and ending, respectively. Literate uses of language related to narrative competence are measured by examining children's use of abstract language and their ability to maintain cohesion throughout extended narrative discourse. Abstract language, specifically the use of metacognitive and metalinguistic verbs, is evaluated through the categories of mental states and character development. The mental state category is included to

document children's use of mental state (e.g., metalinguistic and metacognitive) verbs, such as "say" or "think" or "know" within the narrative. The category of character development also involves the use of mental state verbs to develop characters, and additionally provides information as to whether children are able to differentiate between main and supporting characters throughout the narrative and include dialogue by alternating between the third and the first person. Narrative cohesion is evaluated through the categories of referencing and cohesion. Referencing covers referential cohesion, including the appropriate use of pronouns and antecedents and other verbal clarifiers. Cohesion pertains to conjunctive cohesion (e.g., the use of subordinating and coordinating conjunctions) as well as event sequencing, transitions between events, and appropriate emphasis on critical events. Considering these three aspects of narrative competence (macrostructure, abstract language, and cohesion), the sample can be described as possessing the following characteristics.

Narrative Macrostructure. The majority of the sample provided an introduction to their stories, although most were underdeveloped. A well-developed introduction should orient the listener to the time and place of the story as well as introduce the main character and any important supporting characters. As illustrated in Table 4.15, most narratives in the sample provided at least one setting element such as time or place and some mention of characters. Few (n=13) of the narratives evidenced more "mature" introductions by including some detail about the time, place, and character.

Table 4.15

Frequency of Levels of Introduction with Examples

NSS Intro Scores	Frequency of Score	Examples
1	12	A boy jump.
2	27	One day there was a circus.
3	31	One day three kids were playing baseball.
4	7	There's a circus going on and then a boy is trying to see the circus more closer.
5	6	One day the boy went to the circus and then he wanted to see the lion.

Note. In the examples provided, level 1 = no introduction; level 2 introduces the place and time; level 3 introduces time and characters; level 4 introduces place and main character with some character development; level 5 introduces time, place, main character with some development and secondary character.

Regarding conflict resolution, which is crucial to the episodic structure of narratives according to story grammar schema, most narratives in the sample demonstrated underdeveloped conflicts and resolutions. They either provided a conflict with no resolution or an apparent resolution with no clearly stated conflict (receiving a score of 2), or they provided a conflict and resolution that still left considerable gaps in the story, either because the resolution did not resolve the main conflict of the story or because not enough information was given and the resolution or the conflict had to be inferred (receiving a score of 3).

Table 4.16

Frequency of Levels of Conflict Resolution with Criteria

NSS Conflict Resolution Scores	Frequency of Score	Description
1	14	Non-narrative sequence; no discernible conflict.
2	24	Provides either a conflict or a resolution.
3	26	Provides at least one conflict and resolution, however not the most critical conflict/resolution. They may be underdeveloped and need to be inferred.
4	13	All critical conflicts/resolutions are present; may be underdeveloped but can be logically inferred.
5	6	All critical conflicts/resolutions are clearly developed; resolutions are adequate so as not to leave listener hanging.

Since most narratives received a score of either 2 or 3 on conflict resolution, an example of each is provided in order to illustrate they type of performance characteristic of the sample.

Level 2 Conflict Resolution (provides either a conflict or a resolution)

"Once upon a time, there was (um) two kids that were playing. And three that were play (um, um) basketball. And (um, like) then (um, they pl) the boy threw the (um) ball and it hit the window. And (um) the lady that owned the store was mad and XX stuff. Then (um, um, then) the police came and was grabbing his

hand. And his hat fell down. And he was (um, um) running. And the (um) bat was (um) falling off. And the other one that was playing with him was running."

Level 3 Conflict Resolution (provides an underdeveloped conflict and resolution)

"One day a boy (went out went) went (with the um) with the bat and broke the window of (a) a house. And (and and) a kid was gonna tell the police. And the police came and he grabbed him from the hand. And (um, and the, uh) the boy was scared. (Um) then everybody was looking at him because he was (a boy) a little boy. And ((I think that)) nobody (ha) had seen someone that did that."

The story illustrating level 2 performance provides a conflict or a problem statement (underlined), but everything that follows that is a series of reactions. The police coming and grabbing the boy's hand could be considered an attempt at a resolution but it would require a great deal of inference on the part of the listener to understand that event as the resolution of the episode. There are clues that signal that it is not a resolution, but rather the initiation of a reactive sequence. The action of the police officer is progressive ("was grabbing his hand") and thus the action is not complete. Furthermore, if that statement were an attempt at a resolution a listener would expect more emphasis given to that utterance than the ones that follow, which are essentially minor details about boys running and bats and hats falling. Placement of an utterance within an extended discourse is one way to indicate importance, as are a variety of cohesive devices besides "and" (then, so, because) which serve to link utterances in such a way that the relationship between them is clear.

The story representing level 3 also requires some inference on the part of the listener at points in the story to fill in the sequence of events, however there is enough information to signal both the conflict (a boy broke the window of a house) and resolution of the episode (that the police grabbed him by the hand). Although the listener

would expect more information to follow the police grabbing the boy by the hand, the evaluative statement that follows (and everyone was looking at him because he was a little boy and I think that nobody had seen someone that did that) may signal the end of the story allowing the listener to infer that the resolution was indeed provided. The story illustrated as a level 2 conflict/resolution demonstrates certain qualities that are not uncharacteristic of the English oral narratives of SS-ELLs. McCabe and Bliss (2003) describe frequent use of the past progressive in some Spanish American cultures, as well as an emphasis on maintaining conversational flow over topic maintenance. The latter may result in narratives where extraneous details are tacked on at the end of an otherwise concluded episode, a trait that was commonly observed in this sample. If cultural patterns are thus taken into account, the two examples provided above reflect similar narrative skills, while highlighting the strengths present in the level 3 example and the opportunities for explicit instruction to address the more "immature" behaviors present in the level 2 example.

The category of conclusion overlaps somewhat with conflict resolution in that providing a resolution is part of providing a conclusion, but mature behavior in this domain is marked by providing not only a resolution, but also an ending statement, such as "and they lived happily ever after." Minimal/immature performance (receiving a score of 1) is indicated whenever the child just stops narrating, usually causing the examiner to inquire: "Is that all… are you finished?"

Table 4.17

Frequencies of Levels of Conclusion with Criteria

NSS Conclusion Scores	Frequency of Score	Description
1	25	Stops narrating.
2	17	Signals end of narration ("and that's it" or "that is all").
3	25	A specific event is concluded (e.g., resolved) but no general statement is made concluding the whole story.
4	8	Story is clearly wrapped up using general concluding statements but a significant event or outcome remains unresolved.
5	8	Story is clearly wrapped up using general concluding statements such as "and they were together again happy as could be." No significant event or outcome is left unresolved.

Half of the sample (n=42) either provides no conclusion or minimally signals the end of the storytelling task by saying in effect, "that's all." Another third of the sample (n=25) provides no concluding remarks but does provide a resolution. As alluded to earlier, one function of both introductions and conclusions is that they serve to anchor the events that occur in between them, facilitating cohesive sequencing that aids the listener in identifying the relative importance of utterances. For example, a clear ending compensates to some extent for an unclear resolution in that it allows the listener to infer that the somewhat incomplete resolution provided just prior to the ending was indeed meant to be taken as complete. This was the case with the level 3 conflict resolution

example discussed earlier. Strategies for developing introductions and conclusions in oral and written discourse are routinely taught in grade school and the NSS potentially provides a classroom teacher with information regarding which of those skills needs to be taught individual students and also how the quality of students' narratives changes in response to explicit instruction and practice. When it comes to narrative organization according to the expectations in school settings of what constitutes a well-formed narrative, introductions and conclusions may very well be the 'low hanging fruit.' Simply providing SS-ELLs with a few concrete strategies in these areas may do much to immediately improve their academic oral and written discourse.

Abstract Language. The NSS evaluates children's use of abstract language through the categories of mental state words and character development. The two categories are interrelated, however each highlights slightly different criteria. Levels of proficiency according to the mental states category are contingent upon the quantity and variety of mental state words used in the story to develop characters. Mental state words include adjectives that describe characters' internal states (scared, excited, hopeful, sad, etc.) as well as verbs that indicate their thought processes (know, think) and language (say, tell, etc.). Character development is enhanced by the use of mental state words, however proficient character development also requires that narrators give adequate emphasis to main characters so that listeners can distinguish them from less important characters. Emphasis may be accomplished by providing detailed description of main characters and/or by dedicating a number of utterances to them. Frequencies of performance levels for each category, as well as descriptive criteria, are reported in Tables 4.18 and 4.19.

Table 4.18

Frequencies of Levels of Mental States with Criteria

NSS Mental States Scores	Frequency of Score	Description
1	26	No mental state words.
2	21	A singular mention of a mental state word.
3	16	Use of the same mental state word multiple times OR a singular mention of a mental state word with a reason (he was scared because the lion was chasing him).
4	12	A variety of mental state words used but without reasons.
5	8	A variety of mental state words used with reasons (clearly marked, not implied).

Table 4.19

Frequencies of Levels of Character Development with Criteria

NSS Character Development Scores	Frequency of Score	Description
1	1	There are no characters or characters cannot be determined (e.g., only pronouns or collective nouns are used).
2	21	Characters are present, but no main character stands out.
3	25	Both main and supporting characters are mentioned and there is enough information to distinguish a main character (although very little character development is provided).
4	28	There is an attempt to develop main character using mental state words and/or dedicating a number of utterances to that character. The presence of even one mental state word used to develop a main character automatically qualifies as a 4.
5	8	Main characters and supporting characters are established; main characters are introduced with some description and detail; main characters are emphasized throughout the story; the first person narrative may be used.

For most of the narratives in the sample (n=61), it was possible to identify a main character. Typically this was the boy in either picture, and in some cases the lion. Most narratives (n=57) also included at least a singular mention of a mental state word. While character development and mental states can occur at any point and throughout the

narrative, they greatly enhance the introduction. An example provided earlier and repeated below illustrates the ways one child used mental states, with notable sophistication, to construct an introduction that is highly engaging, creating suspense and anticipation. Unfortunately the child was not able to maintain cohesiveness beyond the introduction and thus the story was rated as a reactive sequence according to story grammar analysis, which offered no way of documenting or describing the child's strengths in the areas of introduction, character development, and mental states.

"There was this kid on a circus XX very happy.

(He was about to get) he was about to get a lot of balloons.

(And he wa) and something was checking him out.

It was a lion, a very angry lion."

The ability to detect variable behavior within a child's performance by documenting specific areas of strength and those needing improvement is an attractive feature of the NSS.

Cohesion. Cohesion is evaluated through the domains of referencing and cohesion. Proficient referencing is characterized by the lack of ambiguity and the effective use of verbal clarifiers to enhance comprehensibility. Proficient cohesion is achieved when events are logically ordered and emphasized appropriately. This is accomplished, in part, through the appropriate use of conjunctions (coordinating and subordinating), which is taken into consideration when evaluating that category. Frequencies of scores and their criteria are reported in Tables 4.20 and 4.21 below.

Table 4.20

Frequencies of Levels of Referencing with Criteria

NSS Referencing Scores	Frequency of Score	Description
1	9	Excessive use of pronouns; no verbal clarifiers used when needed or inappropriate use of articles and other clarifiers.
2	37	Inconsistent (some appropriate, some inappropriate) use of clarifiers such that comprehensibility is compromised and the listener is confused.
3	29	Inconsistent use of referents/antecedents. Some appropriate and some inappropriate referencing. Inappropriate referencing doesn't interfere with comprehension of the basic story.
4	2	No ambiguity in the referencing of characters.
5	6	No ambiguity AND clarifiers (this one, that one, the other one, etc.) are used to enhance comprehensibility.

Table 4.21

Frequencies of Levels of Cohesion with Criteria

NSS Cohesion Scores	Frequency of Score	Description
1	5	A series of disconnected utterances.
2	17	Within the sequence, there is an attempt to connect utterances with active use of cohesive devices.
3	42	Events follow a logical order. Excessive detail or emphasis on minor events may lead listener astray. Equal emphasis on all events because of lack of variety of conjunctions and/or conjunctions are used inappropriately.
4	18	Events follow a logical order (unless violation of order is clearly intentional); some appropriate use of a variety of conjunctions, including subordinating conjunctions.
5	1	Events follow a logical order. Critical events are included while less emphasis is placed on minor events. Transitions between events are smooth.

As briefly discussed in the section describing NAP results, there were considerable problems with referencing in the sample of narratives resulting in ambiguities that diminished discourse coherence. Pronouns lacked appropriate antecedents or references were otherwise unclear. Pronouns, articles, and clarifiers were often used inappropriately or, in the case of clarifiers, not at all. There were some exceptions, however. In the following story, the child's pronoun use was not always

appropriate; however through the ample use of clarifiers the child was able to maintain a fair distinction between the two boys serving as the main characters, with only a couple of ambiguities.

"(Um) there was (a) a kid was with another kid.

They were playing baseball.

(Then) then (they um) the boy (um) throw (um) the ball with (the) the stick and the other boy couldn't catch it.

And it went (in the window) in the window (of another, of another, of anoth, went in the window of another, anoth) X (then on the the the wind the) X.

(Um) so person in store (uh) call the police.

And the police came.

(And) and (un he took) he took the boy.

And then grabbed his hand.

Then the other boy **that he was playing with him**, he was old.

The police couldn't get him.

And then the boy **that the police get**, he throw (um) the stick in the floor.

And everyone was looking at the boys **that they were playing**.

And the lady XX.

And the police came and get the boy.

(Um) then (un, mm) the police grab him.

And took him (uh to his to his mom uh) to his mom's house.

And (um, X police, and uh then) then the boy tells his mom that that other boy (um) didn't catch a ball.

And it went into (the) the window store.

And (um) maybe got a problem too XX."

This story highlights some specific referencing strengths as well as specific syntactical patterns (e.g., the inappropriate inclusion of subject pronoun after the relative pronoun, "that", as in "the other boy that *he* was playing…") that should be targeted for instruction and, if mastered, would greatly improve the referential cohesion and fluency of the story. Even with its syntax errors, however, this story demonstrates (compared with the rest of the sample) a high level of awareness of the listener's needs for clarification in an extended discourse.

In contrast to referencing, cohesion was a relative strength of the narrative sample. The NSS category of cohesion encompasses two categories included in the NAP, those of event sequencing and conjunctive cohesion. Similar to the NAP, the NSS captured the pattern whereby children largely relied on the coordinating conjunctions, "and" and "then," to link utterances in their stories. While the use of some conjunctions are better than none, the overwhelming prevalence of these conjunctions had the effect of leveling the narrative structure such that each utterance is equal in importance to each other utterance. To establish or punctuate varied levels of importance of utterances, it is necessary to link ideas with a variety of conjunctions and cohesive devices that establish both causal and temporal relationships between them. These may include "because," "so," "while," and "when" to name a few. As with others, this performance pattern indicates an opportunity for targeted instruction to increase the repertoire of conjunctions and syntactic structures the child has at his or her disposal. Cohesion is also a category for which the performance of individual students is fairly evenly distributed with half of the sample performing at the emerging proficiency level (score of 3), and roughly 25% of the sample performing above and below that emerging level. Seventy-five percent of the sample achieved a cohesion score of 3 or higher indicating logically ordered utterances. However, overall cohesiveness was somewhat lessened by the presence of tangential or

off-topic utterances. An example of emerging cohesion according to the NSS illustrates the kind of performance that was observed frequently in the sample.

"(um, um, um) one day there was a circus.

(and um, and um, and da) and this was a person.

(and the and) and they hit the lion.

And the lion get mad.

So the lion got out of control.

And then he escape from the wagon.

(and the and the) and the lion want to get to the boy.

And everybody was going home.

(and) and (the) a kid was running.

And the (cl) clown were afraid.

So (he) he let go of the balloons.

(and and the lion) and the lion were really mad.

So he run really fast.

(but) and the kid run fast as he could far away.

(Then the lion fo) and he want to find protection from the lion."

This child demonstrates some strengths with the use of a variety of conjunctions; however, the four-utterance italicized sequence in the middle of the story essentially detracts from the plot line, especially since each of those utterances is given equal emphasis with the utterances that do constitute the plot sequence. Nevertheless, teaching the child to use other cohesive devices such as "meanwhile" to introduce and offset a sub-episode and subordinating devices such as "who" would provide the narrator a strategy for embellishing the details without leading the listener astray. The cohesiveness of the narrative would thus be improved with some minor changes, for example: "~~And~~

meanwhile, everybody was going home and a kid was running. *There was a* clown *who was* afraid so he let go of the balloons."

In summary, each scoring system plus the analysis of narrative microstructure was able to provide information about the characteristics of the sample of SS-ELLs' English oral narratives. While the length of the children's stories varied as a function of picture prompt, the quality of their stories did not. Story grammar analysis revealed that the majority of the stories in the sample did not meet criteria for story grammar analysis because they lacked identifiable goal-directed behavior. Most of the stories were rated as reactive sequences; however, these evidenced considerable qualitative differences upon inspection, which story grammar schema were unable to document or describe. The most salient characteristic to emerge from story grammar analysis was that the stories, by and large, did not include explicit goal-directed behavior on the part of a protagonist or any other character for that matter. Most characters' actions were reactive with the exception of the lion, who frequently "wanted to get the boy." Both the NAP and the NSS provided qualitatively different kinds and amounts of information than did story grammar analysis. Both scoring systems used examiner judgment to rate performance across several categories, some of which were common across measures.

The NAP categorized performance as inappropriate, variable, or appropriate depending on how behavior in a particular domain functioned to increase or reduce discourse coherence. Overall, children's performance was variable to inappropriate. Conjunctive cohesion was a noted exception, evidencing variable to appropriate performance in almost all narratives. Referencing and fluency were particularly problematic areas according to the NAP. There were many gaps in the kinds and amounts of information provided in narratives as well as evidence of patterns that alternated between providing too little or too much of a specific type of information.

Children in the sample tended to provide more evaluation and action than on-topic description.

The NSS evaluated stories according to their macrostructural features as well as qualities of abstract language and cohesion. In terms of macrostructure, introductions and conclusions were minimally developed or omitted altogether in most cases, however there were exceptions. Conflicts and resolutions were also underdeveloped, frequently requiring that the listener inference to fill in missing information. Abstract language use was evident through the presence of mental state words to develop characters. Many stories included singular utterances conveying the mental state of one or another character. Some narratives were more elaborative with mental states and while most character development was minimal but sufficient to distinguish main from supporting characters, some children included dialogue and used the first person to give voice to their characters' thoughts and linguistic expressions. As with the NAP, the NSS also recognized weak and problematic performance in the area of referencing but stronger performance in cohesion (including conjunctions and overall organization of events). While performance patterns could be recognized describing typical performance in each domain, the real strength of the NSS is in its ability to document specific strengths and weaknesses within an individual narrative performance. The NAP is able to do this as well, especially given the modifications that were made while examining this sample of narratives, however the NSS possibly offers a more parsimonious tool (fewer categories) allowing for more nuanced descriptions of narrative performance due to its broader scale.

Stratifying the Sample

To compare the stable characteristics of SS-ELLs' narratives across the scoring protocols used, the sample distribution of the NSS was chosen as a standard by which to identify average, above average, and below average performers. The NSS was chosen because of its relatively normal distribution (see Figure 4.8).

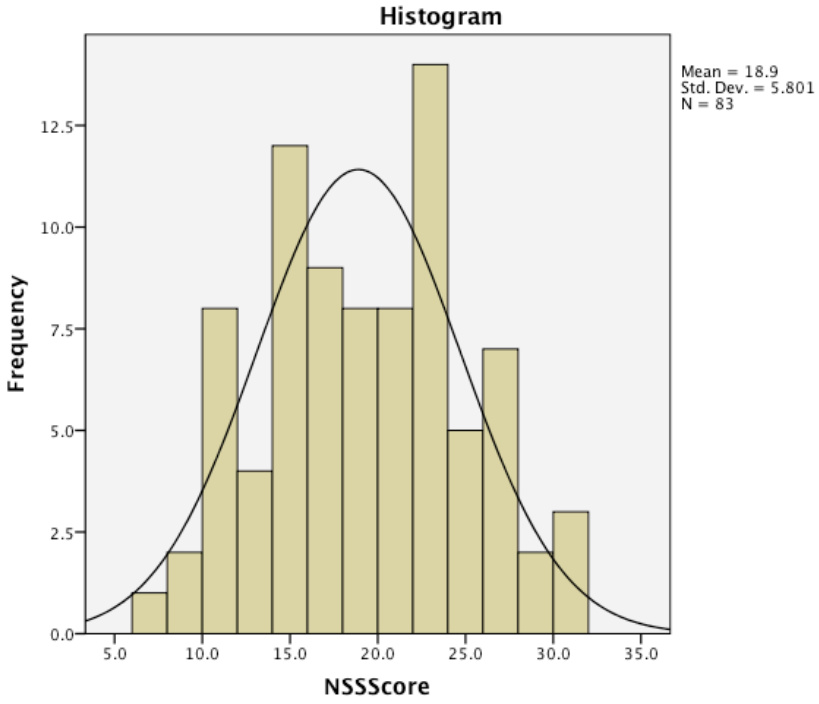


Figure 4.8. Narrative Scoring Scheme sample distribution.

Three groups were created. The average group consisted of those cases whose scores ranged between -1 and +1 standard deviations (=5.801) from the mean of 18.9. After rounding, the range defining average performers was 13-25. Total NSS scores below 13 defined the below average group and total scores greater than 25 defined the above

average group. Variables were computed in SPSS to mark individual cases as belonging to its appropriate group according to NSS distribution and then descriptive reports were generated to identify the story grammar and the NAP scores of the average, below average, and above average cases. Figures 4.9 and 4.10 illustrate cross tabulations of frequency of story grammar and NAP scores, respectively, by NSS group.

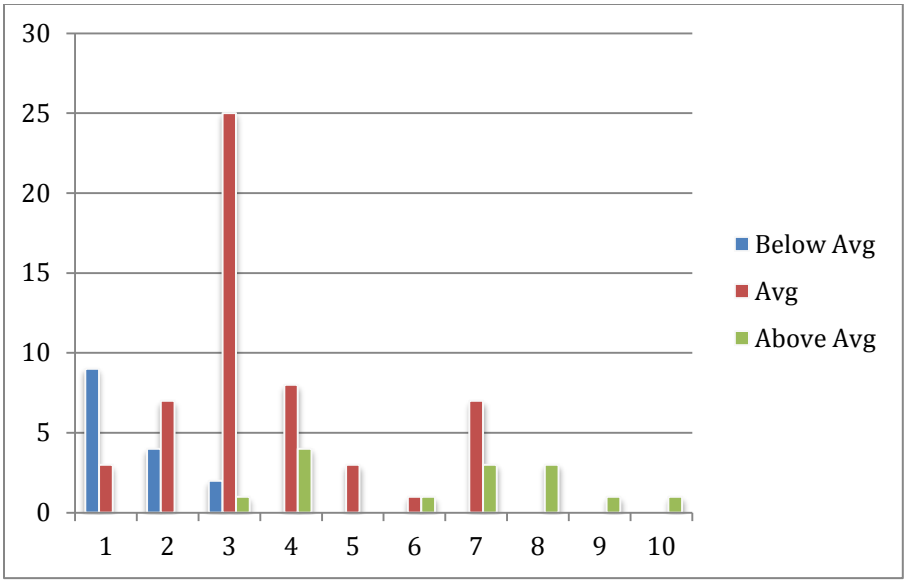


Figure 4.9. Frequency of story grammar scores by NSS group.

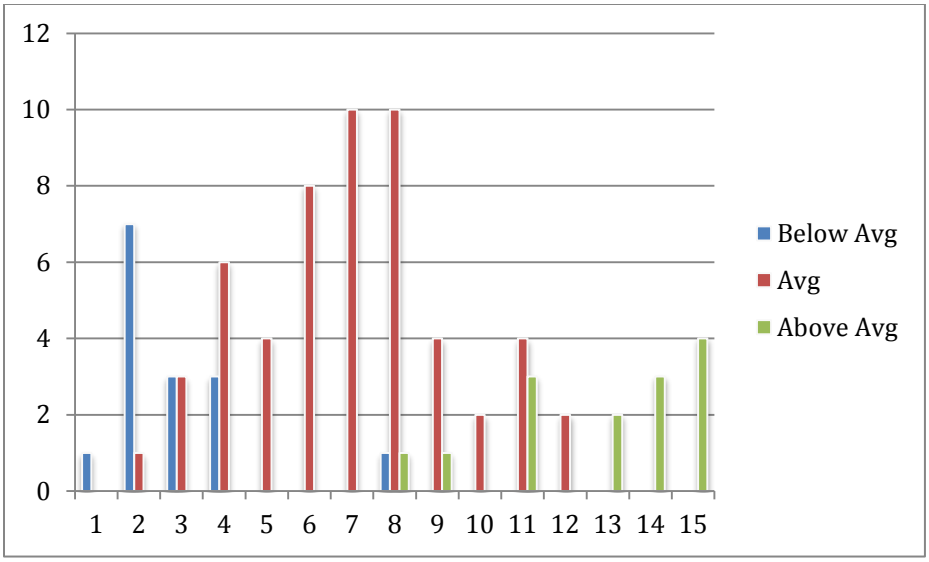


Figure 4.10. Frequency of NAP scores by NSS group.

These figures illustrate that, for story grammar, all of the NSS low scores were also low by story grammar standards, however a considerable number of narratives ranked average by the NSS were clustered at the low end of the story grammar scale and several narratives belonging to the high NSS group were clustered around the middle of the story grammar scale. Likewise, with the NAP, the NSS low group generally received scores at the lowest end of the NAP scale with the exception of one narrative belonging to the NSS low group whose NAP score was 8 out of 16. The NSS average group ranged from a low score of 2 up to an average-high score of 12 on the NAP, however most of the NSS average narratives also received average ratings on the NAP. Both measures additionally identified similar high performers, with the NAP rating the NSS high performers as average to high but mostly high.

The greatest discrepancies in distributions were between story grammar and the NSS. Use of the NAP resulted in a much more similar distribution of ratings, however average performers were still skewed to the left of the midpoint of the scale. If used to

screen children, story grammar and to a lesser extent the NAP would tend to identify more children as low performing than would be expected. Table 4.22 and Figure 4.11 illustrate the numbers of cases categorized as low, average, and high by each system.

Table 4.22

Categorical Distribution of Low, Middle, and High Scores across Measures

	Story Grammar	NAP	NSS
Low	51	28	15
Middle	27	46	54
High	5	9	14

Note. Story grammar low = 1-3; middle = 4-7; high = 8-10. NAP low = 0-4; middle = 5-11; high = 12-16. NSS low = 0-12; middle = 13-25; high = 26-35.

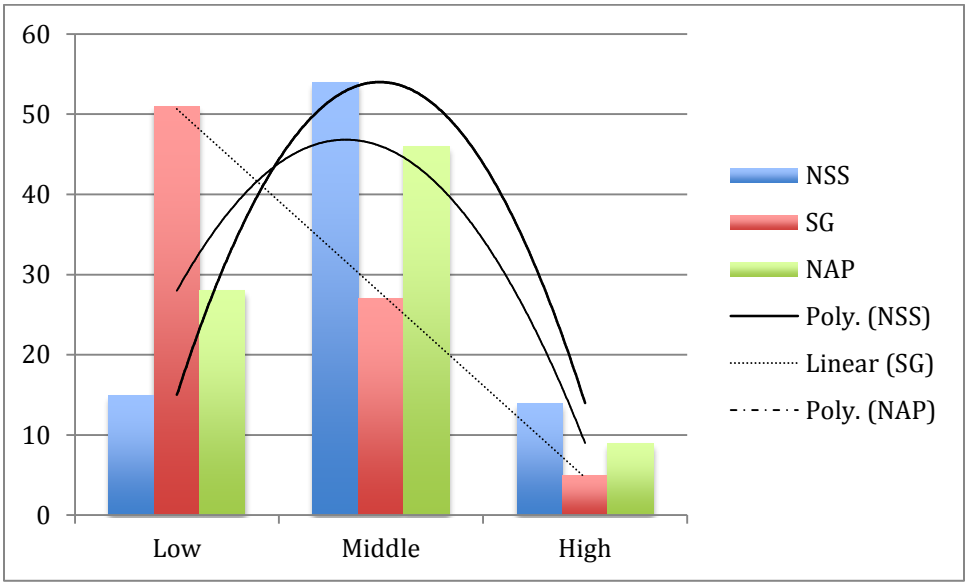


Figure 4.11. Categorical distribution with trendline of low, middle, and high scores across measures.

Cases Rated Consistently Average, Below Average, or Above Average across Scoring Systems

Variables were computed to identify those cases that were: a) rated low by all three measures; b) rated average by all three measures; and c) rated high by all three measures. As a result, 37 cases were identified that had consistent ratings (low, average, high) across all three scoring systems. Of these, 14 cases were identified as low, 18 as average, and 5 cases were identified as high by all three measures. Table 23 identifies those narratives by their study-assigned ID. As all but one of the 42 children in the sample told a narrative on two different occasions (in March/April and in late May of the same year), it was of interest to identify participants for whom both narrative productions received the same consistent ratings across measures. Eight were identified (as indicated by the asterisks in Table 4.23): four categorized as low and 4 categorized as average.

Table 4.23

Narratives Rated Low, Average, or High across Scoring Systems

Cases	Low	Average	High
1	11062_52708	11057_32408*	11101_52808
2	11063_52708	11057_52708*	11109_52708
3	11066_32408	11060_31808	11116_52808
4	11067_31808*	11061_31808*	11117_31808
5	11067_52708*	11061_52708*	11125_40208
6	11069_32408	11064_52708	
7	11102_40208	11065_31808*	
8	11119_32408*	11065_52708*	
9	11119_52808*	11072_52708	
10	11120_40208	11104_52808	
11	11122_31808*	11105_52808	
12	11122_52808*	11108_31808	
13	11123_32408*	11110_52808	
14	11123_52808*	11112_40208	
15		11113_31808*	
16		11113_52808*	
17		11114_52708	
18		11115_40208	
TOTAL	14	18	5

Note. Narrative IDs consist of two pieces of information: the five digit number before the underscore is the participant ID and the five digit number after the underscore is the date the narrative was elicited; * = participants whose two narratives productions were each rated consistently low, average, or high across all three measures.

Select variables were compared to describe the microstructure and macrostructure features of consistently rated below average, average, and above average English oral narrative performance.

Microstructural Features of Low, Average, and Above Average Oral Narrative Performance

Group means were compared for the microstructure variables, number of utterances in the analysis set (NU), number of total words (NTW), number of different words (NDW), type token ratio (TTR), which is an index of lexical diversity calculated by dividing NDW by NTW, and subordination index (SI). Results are reported in Table 4.24.

Table 4.24

Comparison of Means of Microstructure Variables Between Average, Below Average, and Above Average Groups

		Low (n=14)	Average (n=18)	High (n=5)
NU				
	Mean (Std. Error)	5.57 (.894)	13.67 (1.444)	17.6 (1.6)
	Median	4.5	12.5	17
	Minimum	2	7	13
	Maximum	12	27	23
NTW				
	Mean (Std. Error)	31.14 (5.86)	99.56 (13.26)	132.6 (15.17)
	Median	28.5	88.5	136
	Minimum	5	40	87
	Maximum	74	223	174
NDW				
	Mean (Std. Error)	18.36 (2.65)	46.28 (4.86)	61.2 (6.71)
	Median	16.5	43	63
	Minimum	5	19	39
	Maximum	35	96	81
TTR				
	Mean (Std. Error)	.68 (.047)	.50 (.022)	.464 (.016)
	Median	.655	.505	.46
	Minimum	.46	.33	.42
	Maximum	1.0	.71	.52
Subordination Index (SI)				
	Mean (Std. Error)	1.05 (.027)	1.12 (.031)	1.21 (.041)
	Median	1	1.14	1.22
	Minimum	1	.86	1.11
	Maximum	1.29	1.31	1.35

Visual inspection of the group means reported in Table 4.24 suggest that there are differences in the amount of language produced associated with group status. One-way ANOVA confirmed that differences in means between groups were significant ($p < .05$) for all measures (see Table 4.25).

Table 4.25

Between Group Comparison of Means of Microstructural Measures

	Means					
	Total (n=37)	Low (n=14)	Avg (n=18)	High (n=5)	F	Sig
NU	11.135	5.571	13.667	17.6	15.433	**.000**
NTW	78.135	31.143	99.556	132.6	14.194	**.000**
NDW	37.73	18.357	46.278	61.2	16.855	**.000**
TTR	.565	.6843	.5006	.4640	9.813	**.000**
SI	1.1086	1.055	1.121	1.214	3.57	**.039**

Additional comparisons were made to determine whether differences existed between low and average performers and between average and high performers. Differences between low and average performers were significant ($p <.05$) for all measures except the subordination index while those between average and high were not significant for any of the measures (see Tables 4.26 and 4.27).

Table 4.26

Comparison of Means of Microstructural Measures: Low and Average

	Means				
	Total (n=37)	Low (n=14)	Avg (n=18)	F	Sig
NU	11.135	5.571	13.667	19.762	**.000**
NTW	78.135	31.143	99.556	18.405	**.000**
NDW	37.73	18.357	46.278	21.66	**.000**
TTR	.565	.6843	.5006	14.301	**.000**
SI	1.1086	1.055	1.121	2.372	**.**134

Table 4.27

Comparison of Means of Microstructural Measures: Average and High

	Means				
	Total (n=37)	Avg (n=18)	High (n=5)	F	Sig
NU	11.135	13.667	17.6	1.845	.189
NTW	78.135	99.556	132.6	1.536	.229
NDW	37.73	46.278	61.2	2.253	.148
TTR	.565	.5006	.4640	.696	.414
SI	1.1086	1.121	1.214	2.101	.162

These findings can be interpreted to indicate that below average performers produce significantly less language in the oral narrative productions than average and high performers. The amount of language produced does not differentiate average from high performers, however. High performing narratives tended to be longer but not

significantly longer according to Analyses of Variance. However, it could be that the much smaller n size of the high group (n=5 versus n=18 in the average group) rendered differences undetectable at the $p < .05$ level. Subordination Index scores indicate the ratio of all clauses in an utterance (including subordinate clauses) to main clauses. An SI score of 1 indicates that all clauses consisted only of main clauses and a score higher than 1 indicates the presence of subordinate clauses. Higher SI scores indicate greater sentence complexity. There were no differences in sentence complexity between low and average or between average and high groups. Differences in SI scores were only significant between low and high performers.

Macrostructural Features of Low, Average, and Above Average Oral Narrative Performance

To examine macrostructural features that were stable across groups identified as low, average, and high, characteristics according to the NSS are first reported, after which group characteristics according to the NAP will be examined to see if they provide additional information or contrary information to what is provided by the NSS. Tables 4.28, 4.29 and 4.30 report descriptive statistics by group for all NSS scores.

Table 4.28

NSS Macrostructure Performance Characteristics by Group

	Low (n=14)	Average (n=18)	High (n=5)
NSS Intro			
Mean (Std. Error)	1.429 (.137)	2.833 (.146)	4.2 (.374)
Median	1	3	4
Minimum	1	2	3
Maximum	2	4	5
NSS Conf			
Mean (Std. Error)	1.071 (.071)	3.167 (.177)	4.4 (.400)
Median	1	3	5
Minimum	1	2	3
Maximum	2	4	5
NSS Conc			
Mean (Std. Error)	1.429 (.137)	2.944 (.221)	4.6 (.400)
Median	1	3	5
Minimum	1	1	3
Maximum	2	4	5
NSS Total Score			
Mean (Std. Error)	10.286 (.369)	21.5 (.513)	28.8 (1.02)
Median	10	22	29
Minimum	7	17	26
Maximum	12	24	31

Note. Intro = introduction; Conf = conflict resolution; Conc = conclusion; Total Score = composite score, which is the sum of all 7 subcomponent scores.

Table 4.29

NSS Abstract Language Performance Characteristics by Group

NSS Char			
Mean (Std. Error)	1.929 (.071)	3.722 (.109)	4.8 (.200)
Median	2	4	5
Minimum	1	3	4
Maximum	2	4	5
NSS Ment			
Mean (Std. Error)	1.286 (.163)	2.833 (.294)	3.4 (.812)
Median	1	2.5	4
Minimum	1	1	1
Maximum	3	5	5

Note. Char = character development; Ment = mental states.

Table 4.30

NSS Cohesion Performance Characteristics by Group

NSS Ref			
Mean (Std. Error)	1.5 (.139)	2.722 (.177)	3.8 (.374)
Median	1.5	3	4
Minimum	1	2	3
Maximum	2	5	5
NSS Coh			
Mean (Std. Error)	1.643 (.133)	3.278 (.109)	3.6 (.245)
Median	2	3	4
Minimum	1	3	3
Maximum	2	4	4

Note. Ref = referencing; Coh = cohesion.

One characteristic of the low group across all performance categories is that their performance does not appear to vary. In all NSS categories except mental states, low performers achieved a score of either 1 or 2, indicating immature and minimal performance in all categories with no particular strengths. In the mental state category, the range was from 1 to 3, however the median score was 1. By contrast, the average group achieved mean ratings in the emerging proficiency range (between 2.7 and 3.7) but scores reflected a range of performance, indicating variable performance patterns of strengths and weaknesses within individual cases. The high group similarly demonstrated variable performance. To detect differences, ANOVAs were conducted to test the difference in means between low and average and average and high groups. A between groups comparison (low, average, high) of means resulted in significant ($p <$

.000) differences across all components of the NSS. Tables 4.31 and 4.32 report ANOVA results for comparisons between the low and average group and the average and high group, respectively, in order to determine where differences occur.

Table 4.31

Comparison of Means of NSS Components: Low and Average

	Means				
	Total (n=37)	Low (n=14)	Avg (n=18)	F	Sig
NSS Intro	2.219	1.429	2.833	46.956	**.000**
NSS Char	2.938	1.929	3.722	167.426	**.000**
NSS Ment	2.156	1.286	2.833	18.045	**.000**
NSS Ref	2.188	1.5	2.722	26.917	**.000**
NSS Conf	2.25	1.071	3.167	139.615	**.000**
NSS Coh	2.563	1.643	3.278	92.520	**.000**
NSS Conc	2.281	1.429	2.944	29.547	**.000**
NSS Total	16.594	10.286	21.5	282	**.000**

Note. Intro = introduction; Char = character development; Ment = mental states; Ref = referencing; Conf = conflict resolution; Coh = cohesion; Conc = conclusion; NSS Total = composite score.

Table 4.32

Comparison of Means of NSS Components: Average and High

	Means				
	Total (n=37)	Avg (n=18)	High (n=5)	F	Sig
NSS Intro	3.13	2.833	4.2	16.504	**.001**
NSS Char	3.957	3.722	4.8	21.639	**.000**
NSS Ment	2.957	2.833	3.4	.665	**.**424
NSS Ref	2.957	2.722	3.8	7.691	**.011**
NSS Conf	3.435	3.167	4.4	12.886	**.002**
NSS Coh	3.348	3.278	3.6	1.773	.197
NSS Conc	3.304	2.944	4.6	12.413	**.002**
NSS Total	23.087	21.5	28.8	43.229	**.000**

Note. Intro = introduction; Char = character development; Ment = mental states; Ref = referencing; Conf = conflict resolution; Coh = cohesion; Conc = conclusion; NSS Total = composite score.

As with microstructural measures, all differences between low and average group were significant ($p < .000$). Most differences between average and high groups were significant ($p < .05$) as well with two exceptions. Mental states and cohesion did not vary between average and high groups. Therefore, when considering the stable features of performance for each group according to the categories of the NSS, low performers perform significantly lower across all categories. The low performers in this sample did not achieve a rating of emerging proficiency on any of the categories; their performance was considered minimal or immature in all categories. By contrast, variability appears to be a characteristic of both average and high performers, all of whom exhibited a range of performance across categories, although certain performance patterns did distinguish the high from the average performers. Specifically, high performers exhibited what was

considered to be proficient and mature performance in all macrostructural categories (introduction, conflict resolution, and conclusion). Their performance, however, was variable in the domains of abstract language and cohesion. The character development category of abstract language was one in which high performers demonstrated significantly better performance, but not mental states. Likewise in cohesion, the high group demonstrated greater proficiency with referential cohesion as compared with the average group, but not with cohesive devices and the overall organization of stories; in those two areas the groups performed similarly.

Examining the ways that the NAP characterizes this sample of low, average, and high performers reveals similar patterns. Low performers were significantly poorer on all categories as compared with average performers ($p < .000$) and high performers were better in some, but not all, categories as compared with average performers. Table 4.33 reports the differences in means between average and high performers on the NAP.

Table 4.33

Comparison of Means of NAP Components: Average and High

	Means				
	Total (n=23)	Avg (n=18)	High (n=5)	F	Sig
NAP Top	1.652	1.556	2.0	3.652	.070
NAP Ev	1.348	1.278	1.6	1.773	.197
NAP Info PO	1.391	1.222	2.0	15.978	**.001**
NAP Info TCH	1.087	.833	2.0	24.855	**.000**
NAP Info CHEF	.783	.556	1.6	15.881	**.001**
NAP Ref	1.00	.833	1.6	4.128	.055
NAP Conj	1.783	1.722	2.0	1.756	.199
NAP Fluen	.087	.111	.000	.571	.458
NAP Score	9.130	8.111	12.80	29.824	**.000**

The only NAP categories that effectively differentiated average from high performers were the three categories of informativeness. Referential cohesion nearly achieved significance at the $p < .05$ level but did not. Similar to the NSS category of cohesion, event sequencing was not significantly better nor was conjunctive cohesion. Informativeness is where high performers stood out by providing appropriate amounts of information to convey the gist of a story, embellish it with detail to make it engaging to listeners, and by providing all necessary types of information, including description, action, and evaluation.

Story grammar holistic score varied significantly ($p < .05$) among the three groups, with the low group achieving a mean score of 1.429 (between a descriptive and

an action sequence), the average group a mean score of 5.278 (between an incomplete and a complete episode), and the high group a mean score of 8.6 (between a complex and an embedded episode). Since story grammar elements were quantified (only for scores of 4 or greater) by their presence or absence and not evaluated qualitatively, they are not comparable with subcomponent scores of the other measures and thus will not be reported.

SUMMARY

Microstructural analyses revealed that the amount of language with which a child told a story was not necessarily related to its overall quality. While higher quality stories did tend to be longer, it was not the case that the longest stories were the highest quality ones nor that the shortest were the lowest quality. Rather, findings suggest that the quality of narrative orations was independent of their length, suggesting that narrative organization measures may be sensitive to narrative organization skills regardless of the amount of language used to tell a story. Even children with limited English may be able to demonstrate appropriate narrative organization skills when given the task to tell a story about a picture in English. Each scoring system used to evaluate the corpus of narratives included in this study resulted in a different set of characteristics. According to story grammar analysis, most of the narratives would be considered sequences rather than true episodes. Story grammar sequences have been empirically associated with the oral narratives of preschool students whereas goal-directed episodes are expected of students at the second grade level. Thus, story grammar analysis appears to systematically skew the evaluation of narrative performance toward the lower end of the scale, erroneously suggesting that SS-ELLs' narrative skills are less developed than those of their monolingual English-speaking peers.

Story grammar was unable to detect particular strengths in the narratives that were not associated with the goal-directed behavior required to minimally constitute an episode. Both the Narrative Assessment Profile and the Narrative Scoring Scheme, on the other hand, were able to document more nuanced patterns of strengths and weaknesses for the set of narratives. According to the Narrative Assessment Profile, the set of narratives exhibited problems with referencing and fluency as well as sufficient inclusion of certain types of information beyond conveying the basic gist of a story. In other areas, the children's performances were variable, or what might be expected of children their age. The Narrative Scoring Scheme evaluated children's stories by examining several aspects across three broad dimensions: story grammar and story like features; the use of abstract language to develop characters and to provide evaluative commentary; and cohesion, including referential and conjunctive cohesion. Most children were able to demonstrate some relative strengths across these aspects, with the overall sample achieving means indicating mostly emerging proficiencies. In terms of macrostructure, story introductions and conclusions as well as conflicts and resolutions were often underdeveloped, requiring the listener to use inference to fill in gaps in information needed to make sense of the story. Most children were able to use a limited number of conjunctive cohesions effectively with some demonstrating relatively stronger performance in this area. Referential cohesion was problematic for most of the sample.

To identify the stable features of narratives rated consistently low, average, or high by all three systems, the sample was stratified and each narrative was categorized as low, average, or high according to each system. Thirty-seven narratives were identified as those receiving similar ratings by all three systems. Based on this stratified sample, average and above average performers exhibited patterns of relative strengths and weaknesses across the various aspects of the NSS. The differences between average and

above average groups were rarely significant. Where the high group did exhibit differential performance was with the quality of narrative macrostructure, but not its microstructural elements. In contrast, the below average performers did not exhibit such patterns of relative strengths. Their performance was low across all narrative aspects, including both micro- and macrostructure variables, and significantly lower than that of the average group.

When applied to individual stories, the revised NSS generated instructionally useful information. It was able to describe strengths as well as identify specific opportunities for improvement across the several aspects of oral narration it measures. The properties that render it useful are its scale and its distribution of related yet distinct aspects of oral narration, which allow meaningful patterns of narrative performance to emerge.

CHAPTER FIVE

Discussion

SUMMARY OF THE STUDY

This study sought to describe the characteristics of a sample of English oral fictional narratives of Spanish-speaking English language learners in the second grade. It also sought to explore how well each of the three narrative scoring systems - story grammar analysis, the Narrative Assessment Profile, and the Narrative Scoring Scheme - evaluated those characteristics and, to identify the criteria that make for a high quality scoring system for evaluating the narratives of the population of school age SS-ELLs. The sample, which consisted of 83 transcripts told by 42 students (41 of whom generated two stories), was prepared for microstructural (language productivity) analyses by segmenting transcripts into C-units and coding for utterance breaks, mazes, errors, and subordinate clauses. Segmented and coded transcripts were analyzed using the Systematic Analysis of Language Transcripts (SALT) software, which generated reports on various microstructural linguistic measures, such as number of words, number of utterances, mean length of utterance in words, number of mazes, number of utterances with errors, and the subordination index. Data from individual transcript reports were compiled in a dataset, to which were added the results of macrostructural analyses by the three scoring systems investigated in this study. At the start of each coding process, beginning with transcript segmentation, a subsample of 20% of the narratives (n=17) was subjected to a reliability process whereby a second rater and the principal investigator tested and refined the coding methods until a reliable method for coding resulted as determined by acceptable levels of agreement measured by Krippendorff's alpha (see Chapter 3 for details). The reliability process resulted in modifications that improved the

reliability of each scoring system and these modified systems were applied to the remaining 66 narratives, which were coded and analyzed solely by the principal investigator.

Once all narrative analyses were complete, the dataset was imported into SPSS and each of the study's research questions was addressed by exploring patterns and trends in the data. Initial comparisons of means indicated that there were differences in the amount, but not the quality nor the narrative organization, of language produced in stories generated in response to each picture prompt. The characteristics of the narrative sample as a whole according to each scoring system's results were reported, after which the sample was stratified by performance on the NSS, the instrument that resulted in the most normal distribution of scores across the sample. Average, below average, and above average narrative performances were identified and compared with similar categorical ratings of each narrative's performance according to story grammar analysis and the NAP. Three subsets of narratives were identified, consisting of those narratives whose performance was rated (e.g., below average or low performing, average or middle performing, and above average or high performing) similarly by all three scoring systems. These subsets of narratives were then explored to identify the stable features that characterize each level of performance (average, low, and high). Once these features were identified and described, scoring systems were scrutinized to identify those features that were well suited and those that were ill suited to evaluating the stable features of SS-ELLs' narratives. Findings informed a discussion of criteria necessary for a high quality scoring system of SS-ELLs' English oral fictional narratives.

In this chapter, findings regarding the features of SS-ELLs' narratives are discussed relative to what others have found and reported about their oral narrative performance. After that, literature specifically addressing the use of the NSS to evaluate

the oral narrative skills of SS-ELLs is examined to determine whether the findings from this study corroborate others' findings regarding the instrument's reliability and how it characterizes the English oral narrative performances of SS-ELLs. To this end, additional sources of data available to this investigation are presented and explored to further describe the potential usefulness of a modified version of the NSS as a tool for those wishing to evaluate the English oral narrative skills of SS-ELLs. Specifically, NSS means are compared with those of a comparison group of 150 Spanish-English bilingual second grade students in Texas and California whose NSS scores are made available in a database included with the SALT software package and published as part of a different study (Francis, Carlson, Fletcher, Foorman, Goldenberg, & Vaughn et al., 2005). Finally, implications for the assessment of the English oral narrative skills of Spanish-speaking English language learners are discussed along with limitations of this study and recommendations for future research.

THE ENGLISH ORAL NARRATIVE PERFORMANCE OF ELEMENTARY AGE SPANISH-SPEAKING ENGLISH LANGUAGE LEARNERS

One major purpose of this study was to identify a set of characteristics that describe the English oral fictional narratives of second grade SS-ELLs. I employed both micro- and macrostructure analyses toward this end and relied on triangulation of three different sources of the evaluation of narrative macrostructure to identify groups of average, below average, and above average performers. The findings that resulted are now summarized and discussed, first in terms of productivity and microstructural characteristics, and then in terms of narrative organization and literate language features.

The average length of stories in words and utterances was significantly greater for stories elicited by the circus picture, presumably because children had more language and/or knowledge or experience related to the content of the circus picture. Nevertheless,

there was much variability in length of story for both pictures. This finding was expected due to the open nature of the prompt and the minimal contextual support a static picture provides. Others have reported task effects when comparing spontaneous narratives generated by static picture prompts to those generated by wordless picture books (a series of pictures) and also story recalls (Fiestas & Peña, 2004; Pearce, 2003). Of these three types of prompts, static pictures offer the least support, providing only a single snapshot of action from which the child needs to derive a beginning, or what events led up to the depicted scene, and a conclusion, or what events followed and how they ended.

The failure to develop a beginning and end is essentially what constitutes production of a descriptive or action sequence whereby the child labels and describes what is seen in the picture but does not attempt to narrate a meaningful episode. To develop a coherent beginning and ending, however, when no stimuli are provided requires the storyteller to draw upon his or her knowledge of story schema, which is precisely what narrative organization measures including the ones used in this study aim to measure. Thus a static picture prompt is less constraining of a child's narrative productions than are other types of prompts, but it also offers less support for the storytelling task (Hedberg & Westby, 1993; Hughes et al., 1997; Pearce, 2003).

Constraint can be desirable when a study aims to compare narrative productions across children or groups of children. Those interested in describing differences in spontaneous narratives between groups and/or within groups over time most often use wordless picture books as prompts, especially books that have been used extensively for that purpose by other researchers and for which much empirical data have been published. Story length and lexical diversity are known to increase with age in both monolingual and bilingual populations based on narrative productions elicited from wordless picture books (Bedore et al, 2010; Miller et al., 2006; Muñoz et al., 2003;

Uccelli & Páez, 2007). Findings reported here, however, are not comparable because there is no criterion or expected performance level for the length nor for the content of spontaneous stories generated by a static picture.

Nevertheless, the value of describing language productivity outcomes resides in the finding that, while there were differences in productivity or quantity of language associated with picture prompt, presumably because the children in the sample had more content vocabulary and knowledge associated with the circus picture, there were no differences in the *quality* of production associated with either picture. In other words, children were able to tell a story equally well regardless of the amount of language they were able to produce. This demonstrates that the narrative organization skills of SS-ELLs are measurable to some extent independent of surface language skills and that even brief samples of oral narrative language can provide important information about the oral narrative skills of SS-ELLs in the elementary grades (Heilmann et al., 2010; Miller et al., 2006; Montanari, 2004). This is especially important because limited language production in the narratives of SS-ELLs may be mistakenly likened to the impoverished language characteristic of the narratives of monolingual children with language and/or learning disabilities, thus leading to erroneous conclusions regarding the presence of a language or a learning disability in SS-ELLs (Gutiérrez-Clellen et al., 2008; Ortiz, 1997; Roth & Spekman, 1986).

This study shows that some SS-ELLs with limited English proficiency are able to organize a cohesive, well-constructed story even with scant English linguistic resources. It appears to be the case that for SS-ELLs with limited English, narrative organization measures are able to detect both strengths and deficits in narrative organization independently of surface language strengths and deficits. Oral narrative language samples therefore provide teachers of SS-ELLs an opportunity to simultaneously assess

the two interrelated yet distinct domains of micro- and macrostructural language skills to look for patterns that provide insight into a child's abilities to compose a coherent narrative even where limited English language competence constrains the length of narrative production.

The micro- and macrostructural analysis of transcripts in this study resulted in performance patterns that may be understood to characterize typical performance of second grade SS-ELLs. Based on a stratified subsample, one set of patterns characterized the performance of average and above average performers and a different set of patterns emerged specific to only one group of performers, those who were consistently rated as low or below average on each of the measures. The primary performance pattern characterizing the average and above average narratives in the stratified sample was that of variable levels of performance across the various aspects of narratives. Thus children whose performances were average or above average had uneven profiles of strengths. They showed emerging strengths in some areas and more highly developed competencies in others as represented by a range of scores across variables, the particular compositions of which varied from child to child. Low performers, on the other hand, did not exhibit a range of performance on any of the narrative aspects. For all aspects except mental states, for which NSS scores ranged from 1-3, their scores varied only between 1 (minimal or immature) and 2 (minimal/emerging). They did not demonstrate the competencies more typical of their peers on any narrative aspect. Further, the differences between low and average performers were consistently significant ($p < .05$) for both microstructural and macrostructural measures. In contrast, the differences between average and high performers were significant for only some aspects. The narratives of average and of high performers showed no significant differences on measures of length, lexical diversity, complexity, cohesion or abstract language. They mainly differed in the

sophistication of their actual stories – their complexity, coherence, and embellishment – as measured by the NSS narrative organization aspects of introduction, conflict/resolution, and conclusion. High performers additionally outperformed average performers in character development and referencing, two mutually reinforcing categories that contribute to a narrative's overall coherence as well as its sophistication or the qualities that make it engaging to the listener.

Likewise, results of scoring using the NAP reveal that, compared with average performers, high performers include significantly more information of all three types evaluated by the NAP (gist information, detail, and the inclusion of description, action, and evaluation). Therefore, both average and high performers in the 2nd grade can be said to exhibit emerging competencies across aspects of narrative performance while exhibiting relative strengths and weaknesses in specific areas. The characteristics that separate high from average performers appear to be matters of degree with respect to the macrostructural features and content of narrative performance (Muñoz et al., 2003; Uccelli & Paez, 2007). The implications of this are that narrative language samples can help classroom teachers distinguish between students whose surface language performance (e.g., grammaticality, fluency, lexical diversity) is similar but whose literate language performance sets them apart. Narrative samples may therefore help teachers identify appropriate instructional goals based on the assessment of relative strengths and opportunities for improved narrative performance (Gutierrez-Clellen & Quinn, 1993; Muñoz et al., 2003; Ortiz, 1997).

In contrast to the performance characteristics of both average and high performers, the characteristics that define low performers are not simply matters of degree. They do not exhibit the fluctuating and uneven patterns of strengths that both high and average performers exhibit. Rather they are consistently low across categories.

Not only are they low across categories, both their microstructural and macrostructural performance aspects are significantly ($p < .05$) lower than those that characterize average performances. Their stories are actually much shorter in words and utterances and demonstrate much less lexical diversity than those of average performers. Thus, while amount of language does not appear to constrain the formation of good stories as evidenced by the observed patterns associated with picture prompt, one stable feature of low performers is that their narrative productions generate significantly less language. In other words, children with average and above average narrative skills are able to demonstrate emerging levels of performance and a profile of relative strengths with varying amounts of language. The quantity of language they use does not appear to be a factor in the quality of their performances. Low performers, however, exhibit low levels of quality with no apparent relative strengths, *and* their performances are linguistically sparse.

This leads to a very important consideration once a low performer has been identified. More information is needed about the child's English language proficiency as well as his or her oral narrative competence in Spanish (Ortiz, 1997). Without such information, it is impossible to rule out limited language as the cause of poor narrative performance and it is also impossible to make any conclusions about the child's true narrative capabilities. While it has been established that narrative quality can be demonstrated independently of the quantity of language produced, there is certainly a threshold at which a lack of vocabulary and communicative proficiency with English makes it impossible for a child to produce a narrative in that language, regardless of their metalinguistic capabilities (Bedore & Peña, 2008; Montanari, 2004). This is an issue of performance, not ability. So it would be essential to give such a child an opportunity to perform in his or her native language in order to determine whether the low performance

pattern is present in that language as well. If it is, this may suggest the need for instructional intervention, ideally in the native language, and the monitoring of the child's responsiveness to intervention as a next step (Gutierrez-Clellen & Quinn, 1993; Ortiz, 1997; Peña et al., 2006).

The features described of low performers are similar to the characteristics that have been described of both monolingual and bilingual children with language and/or learning disabilities (Bedore & Peña, 2008; Boudreau, 2008; Cleave et al., 2010; Roth & Spekman, 1986; Roth, Spekman, & Fye, 1995). Specifically, the narratives of children with language and learning disabilities are developmentally lower than those of their peers. They tend to be sparse, less complete, and less complex (Bedore & Peña, 2008; Roth & Spekman, 1986). They are also less coherent and they fail to provide adequate information to meet the comprehension needs of the listener, and so they are difficult to make sense of. Given adequate language with which to compose a narrative, the English oral narrative language samples of SS-ELLs offer valuable insight into their narrative skills. While stylistic differences between narrative productions across languages have been noted, the quality of spontaneous narrative organization is remarkably similar across the two languages of an English language learner and therefore the samples collected in English are reliable indicators of oral narrative proficiency (Bedore & Peña, 2008; Fiestas & Peña, 2004; Gutiérrez-Clellen, 2002). The important difference between monolingual and bilingual children, however, is that for SS-ELLs who exhibit these characteristics in their English oral narrative performances, no conclusions can be drawn without additional information, including performance assessment in the child's native language.

THE NARRATIVE SCORING SCHEME AS A MEASURE OF THE ORAL NARRATIVE SKILLS OF ELEMENTARY AGE SPANISH-SPEAKING ENGLISH LANGUAGE LEARNERS

The Narrative Scoring Scheme is a developmentally sensitive measure of the oral narrative organization skills of monolingual and bilingual children (Heilmann, Miller, & Nockerts, 2010; Miller et al., 2006). Compared with story grammar analysis and the Narrative Assessment Profile, the NSS produced a more normal distribution of scores, identifying proportionally more average performers, many of whom were rated low or low-average by story grammar and by the NAP. Further, it has been shown to be a reliable measure of oral narrative macrostructure when raters are trained in its use (Miller, 2006). Miller and colleagues achieved a Krippendorff's alpha level of .74 for coding agreement on the English transcripts of their Spanish-English bilingual participants. Although this is an acceptable alpha level for drawing tentative conclusions (Krippendorff, 2013), I sought a minimal alpha of .80, given which coder agreement may be considered reliable. To minimally achieve this desired alpha, it was necessary to modify the original rubric to include descriptions for what criteria constitute scores of 2 and 4, which the original rubric left undefined. Further, it was necessary to develop more detailed descriptions of the anchor scores of 1, 3, and 5 as well. Critics of the NSS have also noted concerns about the brevity of its description of scoring criteria (Peterson, Gillam & Gillam, 2008). Only after defining these levels were we able to achieve the desired alpha levels for NSS coding agreement. The NSS has been empirically tested in studies with typically developing monolingual and Spanish-English bilingual elementary age children across the United States. A byproduct of some of this research, there exists a dataset[1] (SALT dataset) with which to compare the NSS results from this study.

[1] Language samples were collected and transcribed as part of the grants HD39521 "Oracy/Literacy Development of Spanish-speaking Children" and R305U010001 "Biological and Behavioral Variation in the Language Development of Spanish-speaking Children", funded by the NICHD and IES, David Francis, P.I., Aquiles Iglesias, Co-P.I., and Jon Miller, Co-P.I.

Language samples in the SALT dataset were collected from SS-ELLs in second grade classrooms in Texas and California. Children produced a unique story retell in English generated in response to the wordless picture book, "One Frog Too Many," by Mercer Mayer. Examiners first modeled the story in their own words with the aid of a script, and then left the book with the child, moved slightly away, and asked the child to retell the story. Narrative productions were recorded, transcribed, and coded for microstructural analysis in SALT and macrostructural analysis using the NSS. The second grade sample consists of 150 stories. Table 5.1 provides a comparison of means of NSS scores between this study's 83 narrative performances and the 150 included in the aforementioned dataset, which is included with the SALT software package. Unpooled t-tests were used to compare means for each narrative aspect. Differences were significant ($p < .05$) for three NSS subcomponents (mental states, reference, and conclusion) as well as the NSS composite score.

Table 5.1

Mean NSS Performance of Sample Compared with Performance of Spanish-speaking ELLs Included in SALT Database

	Study Sample				SALT Database Sample			
	N	Range	Mean	St. Dev.	N	Range	Mean	St. Dev.
Intro	83	1-5	2.614	1.069	150	0-5	2.76	.85
Char. Dev.	83	1-5	3.253	.9858	150	1-5	3.05	.81
Mental *	83	1-5	2.458	1.328	150	1-5	3.3	1.05
Reference *	83	1-5	2.506	.9800	150	1-5	2.81	.85
Conflict Res.	83	1-5	2.675	1.149	150	0-5	2.81	.88
Cohesion	83	1-5	2.916	.8440	150	0-5	2.99	.88
Conclusion *	83	1-5	2.482	1.282	150	0-5	2.99	.87
NSS Score *	83	7-31	18.90	5.801	150	6-35	20.84	4.82

Note. Means were compared using an unpooled t-test. * = differences are significant ($p < .05$).

Figure 5.1 provides a visual comparison of the subcomponent scores reported in Table 5.1. Although some differences in means were significant, performance between the two groups appears to be comparable.

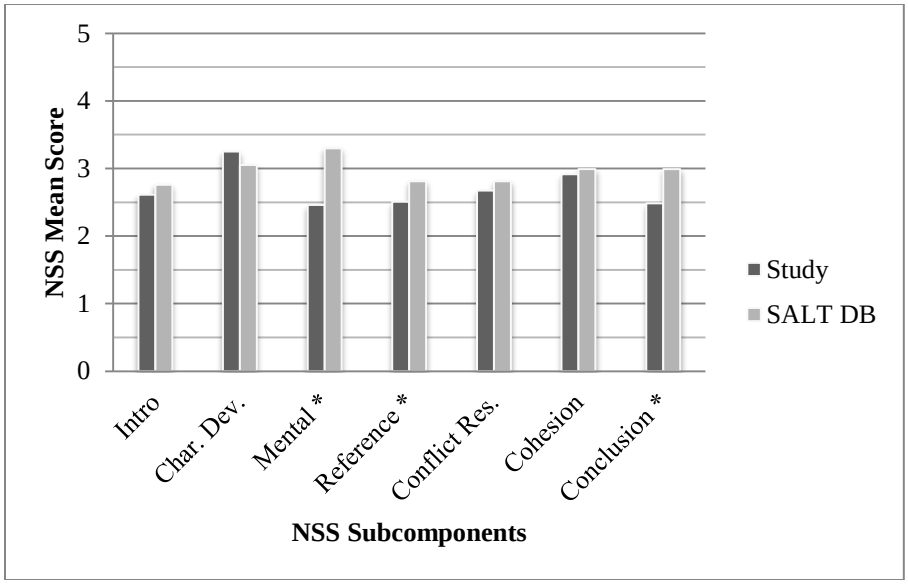


Figure 5.1. Comparison of mean NSS subcomponent scores between study sample and SALT database sample. Means were compared using an unpooled t-test. * = differences are significant ($p < .05$).

It is interesting that the domain of abstract language, which consists of the subcomponents of character development and mental states, was a relative strength across subcomponents for each group. However, the SALT database group had higher means for mental states while this study's sample exhibited slightly but not significantly higher means for character development. This pattern could quite possibly reflect the presence of more mental state words included in the scripts read to the comparison group prior to story retell. This study's sample, in contrast, was free to include or not include mental state words at each child's discretion; no mental state words were modeled ahead of time. Due to the much greater contextual support provided the comparison group for their story generation, one would expect their stories to result in higher NSS scores, which they did.

Both groups were able to generate stories with similar macrostructural qualities in terms of introductions and conflict resolutions. It is not surprising that the comparison group generated better conclusions given that they retold a story from a script that

probably included a conclusion, which would have been highly retrievable during the retell because it is the last part of the story heard. The subcomponent measures of cohesion and reference reflect microstructural performance. Both groups produced comparably cohesive stories as measured by the variety of conjunctions appropriately used and the logical sequencing of events. The comparison group performed better in the area of referential cohesion. This, again, may reflect the story language modeled by the examiner and/or the fact that the examiner moved slightly away from the child prior to the retell. This removed the possibility of shared access to the pictures requiring references to be made explicit during the story retell. This study's sample, on the other hand, generated a story from a picture that sat between the child and the examiner. It may be the case that children assumed that the examiner shared knowledge of the set of characters about whom the story was being told because those characters were depicted and were visually accessible to both the child and the examiner. It could also be that the children in our sample used nonverbal means of clarifying references, such as pointing. This task effect is certainly a possible factor in the relatively low scores for referential cohesion achieved by the children whose 83 narratives were analyzed in this study.

Another possible explanation for differences in mean scores between the two groups is that the sample of 150 bilingual 2nd grade students composing the comparison set were described as "typically developing," as determined by normal progress in school and the absence of special education services. The sample of forty-two participants included in this study represented an entire bilingual 2nd grade cohort at the participating school and may have included students with disabilities. I therefore removed the low performers' scores from the calculation of means to compare means of the subset of participants from this study assumed to be "typically developing" (e.g., those in the high

and average groups) due to their performance patterns. The recalculated mean NSS scores are reported in Table 5.2.

Table 5.2

Mean NSS Performance of Average & High Performers (n=68) Compared with Performance of Spanish-speaking ELLs Included in SALT Database

	Study Sample Avg & Hi Perf.				SALT Database Sample			
	N	Range	Mean	St. Dev.	N	Range	Mean	St. Dev.
Intro	68	1-5	2.882	0.9701	150	0-5	2.76	.85
Char. Dev.*	68	2-5	3.529	0.8547	150	1-5	3.05	.81
Mental *	68	1-5	2.721	1.3026	150	1-5	3.3	1.05
Reference	68	1-5	2.721	0.9279	150	1-5	2.81	.85
Conflict Res.	68	1-5	3.015	0.9696	150	0-5	2.81	.88
Cohesion	68	2-5	3.191	0.6291	150	0-5	2.99	.88
Conclusion	68	1-5	2.721	1.2795	150	0-5	2.99	.87
NSS Score	68	14-31	20.779	4.5837	150	6-35	20.84	4.82

Note. Means were compared using an unpooled t-test. * = differences are significant ($p < .05$).

When low performers, for whom there may be a question regarding the presence of language or learning disabilities, were removed, the characteristics of each group were much more similar; significant differences were only found in the two aspects of the abstract language domain. However, if an abstract language composite score is calculated by summing the means of character development and mental states, the mean composite abstract language scores are 6.25 for this study's average and high performances and 6.35 for the comparison group's typically developing performances. Combining these scores thus eliminates any differences between the groups. The two

categories have some overlap of criteria but are meant to be sensitive to different aspects of literate abstract language. Another plausible explanation for the differences found here is that the modifications we made to the scoring criteria to increase reliability of the rubric resulted in a slightly different distribution of abstract language performance expectations across the two aspects of mental states and character development. It was relatively harder for our sample to achieve a high score on mental states as compared with character development. A single mention of a mental state would earn a 2, and that same single mention of mental state, if it were used to develop a main character, would result in a score of at least 4 for character development.

These between-group comparisons of NSS scores suggest that findings from the current study regarding the characteristics of 2nd grade SS-ELLs' oral English narratives corroborate others' findings. The few differences that exist in subcomponent scores are not widely divergent and they may be plausibly explained by the differences in the tasks and in scoring criteria previously described. Thus the NSS appears to be a reliable measure for describing the oral English narrative performance patterns of SS-ELLs.

INCONGRUENCE BETWEEN STABLE FEATURES OF SS-ELLS' ORAL ENGLISH NARRATIVES AND PROPERTIES OF NARRATIVE SCORING SYSTEMS

With a sample of 83 narratives generated by 42 SS-ELLs in the second grade, a range of performance is expected. A narrative scoring system needs to be able to differentiate levels of performance in a way that reflects the actual performance patterns of the population. An obvious mismatch is apparent with scoring systems that cluster most scores of a broad sample at one end of the scale or another. This was the case with story grammar analysis, which rated most narratives as sequences and not true narratives. Because story grammar narrowly defines a "story" as an episode possessing goal-directed

behavior, it excludes for evaluation many of the narratives in this sample and only provides information about what most narratives lack, not what they possess.

Story grammar is potentially useful in the context of an instructional intervention where story grammar is taught as a means to develop oral and written narratives, however it still has shortcomings when applied to the narrative styles represented in this study's sample. In this sample, although many characters were mentioned and it was usually possible to pick out a main character, there was generally a de-emphasis of a single protagonist with goal-directed behavior. In many cases, scenes were described in which various characters responded to the problem situation. While this may reflect immature storytelling abilities, it may also reflect a cultural tendency to posit the family unit or the community as the unit of agency rather than a single individual (Berman & Slobin, 1994; Gorman, Fiestas, Peña, & Clark, 2011; Greenfield, Keller, Fuligni, & Maynard, 2003; Gudykunst, Matsumoto, Ting-Toomey; Nishida, Kim, & Heyman, 1996). Therefore a problem situation is interpreted as affecting everyone in the picture and thus both problem and solution do not belong to a single protagonist, but are shared. If this were indeed a cultural pattern and not a deficit in storytelling skills, many of the children in this sample would be viewed as having emerging strengths in this area because they seemed to make a concerted effort to include all of the characters at their disposal in their narration. Because of the distribution of internal responses, reactions, and attempts amongst various characters, it was often hard to identify which character should be considered the main character. Often more than one qualified and a judgment needed to be made. Depending on who was chosen as the main character, the entire story structure could be interpreted quite differently as well, resulting in different holistic scores. Because of these characteristics, unless story grammar in its current conception is specifically taught and children's response to that instruction is being evaluated, story

grammar does not appear to be an appropriate measure of SS-ELLs' English oral narrative skills.

It may be both possible and desirable, however, to modify story grammar analysis such that it would be a more valid measure and better able to capture culturally different narrative styles. To do this one would need to recognize the existence of multiple or distributed protagonists when coding episodes for story grammar elements and for holistic levels. This would essentially entail coding parallel episodic structures centered on the motivated behaviors of each protagonist. The current story grammar coding process is overly constraining of SS-ELLs' narratives due to its linearity. Parallel coding and a redefined notion of protagonist is worth exploring to understand whether story grammar could provide valid and instructionally relevant information in the analysis this population's fictional narratives.

Both the NAP and the NSS are capable of identifying a range of behaviors and describing both strengths and weaknesses in narrative performance. The NAP in its unmodified form is not able to discern strengths and weaknesses very well in certain categories. Given that the three types of informativeness were the only aspects that differentiated average and high performers on the NAP, it would be useful to gather more specific information about children's performance in each of those categories. For example, appropriate performance in the category of informativeness according to the chef was defined as the presence of each ingredient and so variable and inappropriate performance was defined as performance in which one or two of those ingredients were missing. However, this gives us no information about which ingredients were missing or if performance patterns exist whereby one ingredient is emphasized over each of the others in a population. The additional analyses applied here aimed to provide these more specific types of information and found that evaluation, indeed, was more prevalent than

description or action in the subsample that was analyzed. Furthermore, the category of fluency as it was defined was mismatched with the kind of narrative analysis (of transcripts) undertaken in this study. While it is true that nearly all of the narratives in the sample were highly dysfluent, because raters were reading transcripts rather than listening to oral narrative performance, the dysfluencies were less of an issue and did not actually interfere with comprehension of the narrative once they were offset in the transcript segmentation process. Rating the entire sample as dsyfluent tells us very little about their fluency. The designers of the protocol additionally measure fluency in terms of words per minute. Words per minute would add information to the fluency category that would make it more interpretable with respect to this sample.

The NAP was designed to be an interventionist's tool, to provide information useful to identifying aspects of narrative performance that may be targeted for intervention. Further, it is based on the performance of conversational narratives, not fictional ones. It is therefore not perfectly suited to the fictional narratives investigated in this study nor is it well suited to sorting performance within a sample due to its small (3 point) scale. However it does provide some novel information to the NSS, specifically a useful way of describing the content of children's stories in terms of their informativeness.

The NSS appears to be well suited to evaluate the English oral fictional narratives of SS-ELLs because it appears to appropriately identify average, low, and high performers and it is capable of providing very specific information on three important aspects of discourse coherence: overall organization and story-like features, the use of abstract language, and cohesion. The NSS required some revision to the published rubric to make it reliable between two raters with this sample of narratives. Specifically, it was necessary to develop criteria for points that fell between the minimal, emerging, and

proficient places on the continuum. Once those were developed and tested, however, the instrument was reliable and capable of identifying low, average, and high performers and the features that differentiate them.

Low performers scored low, in fact significantly lower than average, across all narrative aspects for which they were measured. Their narrative productions were significantly shorter and were not "true" narratives, in the sense that they were not organized around a central topic that included a conflict and attempts to resolve it. They were not stories; rather they were unrelated or loosely related utterances describing objects or actions in the picture. Importantly, low performers' performance patterns were flat; no aspects of their performance were rated more highly than any other aspects of their performance.

Average and high performers, on the other hand, shared the characteristic that their performances were likely to show relative strengths and weaknesses and the difference between the two groups was mainly one of degree. There were no significant differences between average and low performers in terms of length of stories or amount of language produced. The literate features of language observable in narrative discourse, specifically abstract language and cohesion, were present to varying degrees with varying quality with both the average and high groups. The high group appeared to exhibit somewhat better referential cohesion but not lexical cohesion, and both groups included mental state words variably in their narratives although the high performers were able to use those and other features of abstract language to effectively develop characters. The high group's narrative productions exhibited better story-like features, including those described by story grammar. They included more fully developed introductions and conclusions that frame episodes with complete and mostly explicit conflicts and resolutions.

A high quality narrative scoring system must have certain criteria to sort out the low from the average and high performers and also provide useful information to help teachers develop instructional goals for individual students. The system must be capable of documenting strengths and weaknesses across various aspects of narrative discourse. If a system only looks at one feature, such as story grammar for example, it will be unable to detect the variable patterns of strengths exhibited by the average group, essentially causing most children to be rated low, as was the case with story grammar analysis in this study. If a system's rubric uses too short of a scale, such as the three-point scale used by the NAP, it may tend to lump more children's performance across various aspects in the middle. The information provided by shorter scales also may be insufficient to the needs of the teacher when writing instructional goals for children. They provide less specific information unless they are augmented by qualitative description of narrative performance. It was necessary to add levels of scoring to the NAP used in this study in order to glean from it specific information on the characteristics of the sample's narratives.

A system also needs to be user-friendly and reliable, so it must be parsimonious. The NSS had many desirable features in this respect. It required some modifications to increase reliability and to clearly define points on the five-point scale that represented different levels of proficiency for each narrative aspect. Importantly, it did not rely on any sole category to evaluate proficiency within a domain. Abstract language and cohesion each consisted of two aspects and story-like features consisted of three. These qualities made it adept at documenting the uneven performance patterns of the average and high performers and also made it possible to identify the low performers. Were it to be used by classroom teachers of SS-ELLs, it would provide valuable information to

guide a teacher's assessment of both the strengths and the needs of students with respect to their literacy-related oral language skills.

CRITERIA FOR A HIGH QUALITY INSTRUMENT FOR THE ASSESSMENT OF THE ENGLISH ORAL NARRATIVE PERFORMANCE OF SS-ELLS

A high quality instrument of the oral narrative performance of SS-ELLs needs to first and foremost be able to reliably assess performance across micro- and macrostructural narrative aspects, allowing patterns of strengths and weaknesses within individual cases to emerge. If an instrument narrowly measures only singular aspects of performance (e.g., story grammar or grammaticality) or employs too narrow a scale, patterns will be indiscernible. It is precisely the existence of patterns of relative strengths that marks typical language development in bilingual children (Bedore & Peña, 2008). As more data are published regarding both the microstructural and the macrostructural characteristics of SS-ELLs' oral English narratives, the ranges of typical performance across various aspects and for children of different ages and stages of English language development will be better defined, resulting in the ability to better identify appropriate assessments (Laing & Kamhi, 2003).

Secondly, an instrument needs to provide instructionally useful information, specifically regarding an individual child's needs for instruction informing their instructional goals and objectives; and it needs to be sensitive enough to be able to detect changes over relatively short periods of time. Using a combined narrative measure, Uchikoshi (2005) demonstrated the effectiveness of a narrative skills intervention at increasing the English oral narrative skills of SS-ELLs in kindergarten. The intervention and progress monitoring occurred over the course of a school year. The combined narrative measure, similar in its aspects to the NSS, was able to detect significant growth

at intervals of 12 to 16 weeks. This level of sensitivity is critical to making such an instrument useful to instructional planning and decision-making.

To the extent possible, without making an assessment overly complicated, it ought to triangulate or combine sources of information regarding language performance within a particular domain. In other words, it should avoid measuring any one domain with only a sole aspect or component. The NAP provided a good example with its three aspects of informativeness. Taken alone, any one of those aspects provides important but limited information about the content children include in their stories. Given the three aspects, which were distinct yet interrelated, it was possible to discern more nuanced patterns of strengths providing much more specific information that could guide a teacher in establishing learning objectives for a particular child or group of children. For example, knowing that many of the children in this study's sample tended to produce evaluative statements in their narratives yet failed to provide adequate information regarding outcomes of events provides teachers with specific foci for instructional goals. Likewise, the NSS evaluated seven narrative aspects representing three narrative domains: story grammar, cohesion, and literate or abstract language. Because each domain consisted of two or three distinct yet interrelated aspects, it was possible to recognize relative strengths and weaknesses within and not just between domains.

Finally, a quality narrative assessment, such as the NSS, needs to be user-friendly (e.g., "teacher-friendly") if it is going to be instructionally useful. Most narrative measures have been developed and extensively used by speech pathologists. Speech pathologists have very different professional knowledge and skillsets than do elementary school teachers and thus it is very likely that an instrument such as the NSS could only be used by teachers after appropriate training and opportunities for guided practice in its use. The constructs it measures will need to be taught and operationalized for the benefits of

those who have not been exposed to such knowledge during their professional preparation. Ultimately, it is unlikely that even with such training the instrument would be used by teachers if it is cumbersome, time-consuming, unreliable, or produces information they can obtain in some other more familiar way.

Given the complexity of training required to reach proficiency in scoring and interpretation of narratives, these types of assessments are probably best done by speech pathologists, but in collaboration with bilingual education teachers, particularly if the speech pathologists are not bilingual. The type of data the measures provide are crucial to assuring that ELLs develop academic language skills and that they have an adequate foundation for literacy development.

LIMITATIONS OF THE STUDY AND RECOMMENDATIONS FOR FUTURE RESEARCH

There were several limitations to the current study that need be addressed. First, the data came with very little information about the children who produced the narratives. We know they were in a second grade transitional bilingual program at an urban school in central Texas. We had information regarding gender. Other than that, we knew nothing about their histories or their academic profiles, including literacy levels. Of their levels of English oral language proficiency, we knew only their LAS-O scores, which averaged 2.64 with a median score of 3, as reported in Table 3.1. This tells us that, according to the LAS-O, most students were able to communicate a basic story but with notable errors and dysfluencies, consistent with our findings. We didn't know if any were receiving special education services or speech-language therapy. We knew nothing of their home lives, specifically what languages were dominant in the home and who spoke them. We knew nothing of their immigration generational status or when their immediate family arrived in the U.S. Nor did we know about their schooling background

and how much school they had attended and in which languages they had been instructed. Finally we also did not know about their levels of oral language proficiency and literacy in their native language of Spanish. With all of these missing pieces of information, it was possible only to describe the characteristics of the stories on their own merits, but not to draw any conclusions or hypotheses about the children who produced them. Future studies ought to incorporate variables pertaining to participant characteristics in their design.

Even if we had rich data regarding participant characteristics, the samples analyzed in this study were collected under only one narrative task condition. Even the best narrative measure is likely to produce different results depending on the conditions under which the narrative sample was elicited. Reports of language and task effects in oral narration abound in the literature (Allen, Kertoy, Sherblom, & Pettit, 1994; Fiestas & Peña, 2004; McCabe et al., 2008; Morris-Friehe & Sanger, 1992; Gazella & Stockman, 2003; Pearce et al., 2003; Schneider, 1996; Schneider & Dubé, 2005). With respect to fictional stories, there are many levels of contextual support available to aid narrative production. While a static picture is minimally constraining compared to a story retell task or a story told in response to a sequence of pictures, for various reasons it may not result in a child's optimal narrative performance. The scenes depicted in the prompts that elicited this narrative sample may have been familiar to varying degrees to the children. A child's level of familiarity with what goes on at a circus or in a game of street baseball may greatly impact his or her ability to generate plausible inferences to construct the beginning and the end of the story beyond what is depicted.

There are many considerations when choosing task conditions, some of which have been discussed. In this study, referential cohesion may have been compromised by the fact that the picture was accessible to both the child and the examiner. It is possible

that children used nonverbal gestures such as pointing to refer to the characters about whom they were speaking, or that the limited depiction of characters made it conversationally acceptable for the child to mention "the boy" without first introducing him because, perhaps, there was only one boy depicted and he was the most prominent character in the picture. On a pragmatic level, the child might reasonably assume that no introduction is necessary; the examiner knows who "the boy" is because he or she has access to the same picture. There are many contextual factors that need be considered in choosing a narrative elicitation protocol. Ideally, any conclusions drawn about an SS-ELLs' narrative skills would be based on the outcomes of different types of tasks, including personal and fictional narratives, story generations and recalls, provided with differing levels of support, according to the individual characteristics of the child and what will enable him or her to produce the best sample that they can.

Along with task effects, language effects are also an important consideration. The oral language competencies of elementary age SS-ELLs are distributed between the child's two languages (Bedoré & Peña, 2008). To get a complete picture of oral narrative competency, samples must be taken and analyzed in both languages. A goal for future analysis of this sample of narratives would be the scoring and analysis of the children's Spanish narrative productions. Having access to both Spanish and English narrative productions provides a much more complete snapshot of a child's narrative ability at a particular time given a particular task. This information would make it possible to interpret, for example, the performance patterns of the low performers. If the low performers exhibit similar patterns reflecting impoverished language and narrative organization in Spanish, there would be reason for concern and for further testing. It would also be helpful to test the skills of those children in both languages under conditions offering more contextual support. Finally, the background information

described above would be crucial to the process of problem solving and hypothesizing as to the cause of the low narrative performances in both languages.

The process of becoming reliable using each of the scoring systems necessitated modifications of those systems to better match the characteristics of the English oral narrative samples of SS-ELLs. The narrative examples provided by the developers of the instruments and used to illustrate the application of scoring criteria were typically not at all characteristic of the narratives we encountered and attempted to score. We had to essentially reinterpret the criteria to make them applicable to the sample of narratives. While we were interested in describing some surface language performance characteristics, we had to be careful to not let surface language deficits interfere with our ability to detect strengths in organizational and other story-like features. This required us to operationalize criteria in a much more specific way than the instruments in their original form had done. The resulting modified rubrics, in addition to being reliable, were tailored to the characteristics of the narratives and were thus better able to capture nuances contributing to the observed patterns and the identification of instructional implications. Additional research should address the process of modifying and developing narrative instruments such as these and then testing them widely in a variety of contexts, comparing the information they generate with that generated by other, validated narrative instruments.

Finally, a task for future research entails the development of curricula and media with which to deliver the professional knowledge teachers will need in order to understand and make use of the Narrative Scoring System. Professional development centered on the topics of second language acquisition and the practical application of oral narrative assessments needs to be delivered and evaluated along with specific instruction and practice in the use of appropriate instruments such as the modified Narrative Scoring

System. Further, teachers will need mentoring and coaching as they begin to use the instrument in their classrooms so that they may observe and learn how the information gained from using the NSS can be used to write instructional goals and objectives, to design interventions, and to monitor progress. One recommendation is to encourage collaborative partnerships of SLPs and teachers toward this end. Feasibility studies are needed to evaluate the potential of such practices.

SUMMARY

This study affirmed what we know about SS-ELLs in the elementary grades: that the performance patterns of typically developing bilingual children are variable, reinforcing the need for instruments that examine extended discourse samples holistically while measuring various micro- and macrostructural aspects of language production. The study also found that low performers had different traits: Their performance patterns were more level and low across all aspects. They were significantly different from their peers in all aspects of their narrative productions. This finding is consistent with what we know about the narrative performance patterns of monolingual and bilingual children with LI and/or LD. For the low performers, however, more information is needed to further develop a profile of language competencies across languages and in different communicative contexts.

This study also affirmed that oral narrative samples might be ideal ways for teachers and others to collect meaningful information that will aid them in identifying instructional goals and interventions for the literacy-related language development of SS-ELLs. Additionally, the study confirmed that all scoring systems are not equal to the task and that scoring systems must possess certain features if they are able to accomplish this. The modified NSS has these features, which include the measuring of various micro- and

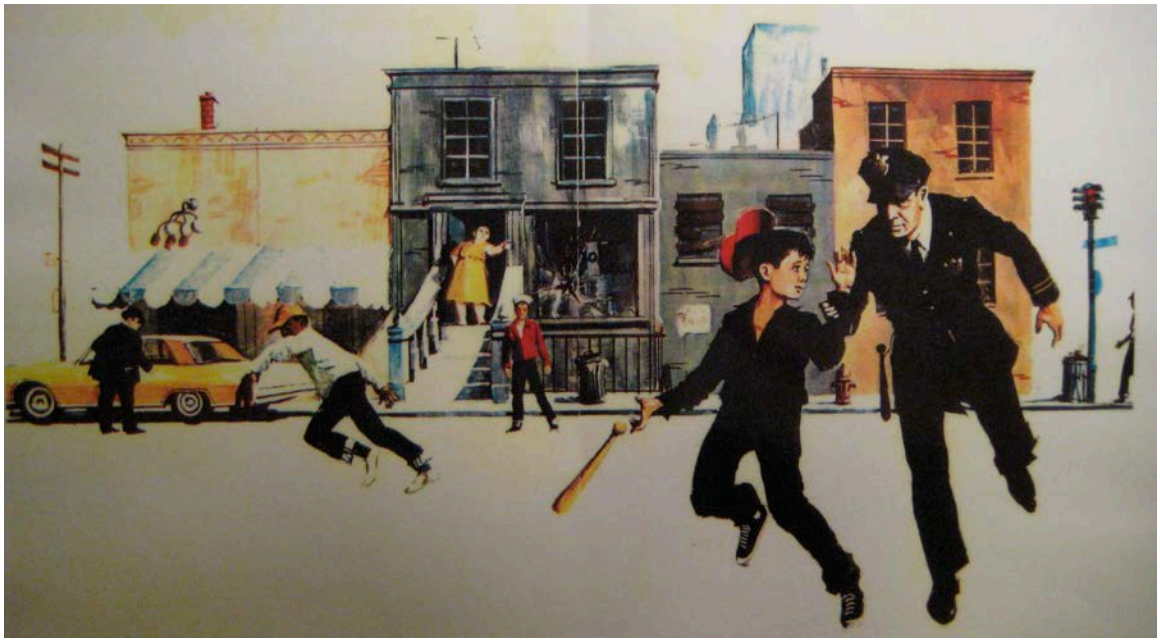
macrostructural domains of oral narration. The oral narrative domains, including story grammar organization, cohesiveness, and the use of abstract language, are evaluated by examining at least two interrelated but distinct aspects. The scale of the NSS is broad enough to capture nuances of performance that effectively make it possible for uneven patterns to emerge and for specific, instructionally useful information to be produced. The main weakness of the NSS is in the lack of specificity of some of its criteria, contributing to difficulties in obtaining reliable agreement between coders. This problem was addressed by modifying the rubric and developing more explicit criteria resulting in the improved reliability and informativeness of the measure.

Finally, the study leaves us with practical considerations for making this tool as user-friendly as possible for teachers, who may have little to limited knowledge of the types of constructs it measures. As well, the findings described herein suggest directions for future research.

CONCLUDING REMARKS

My hopes are that the information reported in this dissertation causes its readers to learn something new and to think differently about the stories generated by elementary age Spanish-speaking ELLs and the usefulness of examining them. I hope it leaves them understanding that the elicitation of oral narratives can and should be integrated into language and literacy instruction and assessment for SS-ELLs and that, given the right tools, they can provide teachers with very specific information regarding children's strengths and opportunities for development and thus they can inform instruction and aid in progress monitoring as well. I hope the reader will deem it useful for teachers to learn more about features of narrative performance and skills and the tools at their disposal to measure or describe them.

Appendix A: Picture Prompts



Appendix B: Scoring Rubrics

APPENDIX B1: STORY GRAMMAR ANALYSIS MODIFIED DECISION GUIDE AND RUBRIC

Story Grammar Analysis Decisions:

1. Read the story.
    a. **Decide → is it appropriate for SG analysis?** In general, **goal-directed behavior must be evident** and thus ***abbreviated episodes*** **are the minimum requirement for a full SG analysis using SG elements.** A *reactive sequence* may receive a limited SG analysis (e.g., it may display one or two SG elements).
    b. *If story is anything less than a reactive sequence,* assign it a story level code but do not attempt to code utterances for SG elements. It may be helpful to categorize the sequence's "***non-narrative story elements***" (from Westby, 1992) as applicable (see examples of these at end of this doc):
        i. Actions (Ac): description of the character's actions
        ii. Internal States (IS): descriptions of character's internal states, such as thoughts, emotions, hunger, sickness
        iii. External States (ES): descriptions of the story environment, such as weather or location
        iv. Natural Occurrences (NO): changes in the environment, such as a violent thunderstorm
    c. **Decide → which goal (if more than one are present) is the superordinate, or most important goal?** Often, stories have goal-directed behavior related to more than one character. If the protagonist is not explicitly stated as part of the setting, it can be difficult to determine which goal is the SUPERORDINATE goal around which the rest of the

story is structured. Stein (1982) suggests "in most story sequences, the initial goal stated or inferred is normally chosen as the most important," however this is not always the case; and "when the initial goal is not used, we have no criteria to guide us in choosing the correct goal. A working theory of importance has to explain how a comprehender's knowledge of human action and motivation influences the decisions made about the importance of an event (p. 323).

d. Assign a Story Level:

i. **Level 1**: Descriptive sequence – describe characters and actions but no causal relationships

ii. **Level 2:** Action sequence – actions described in correct chronological order but no causal relationships; lists actions that are chronologically but not causally ordered

iii. **Level 3:** Reactive sequence – series of actions with some causal relationships (e.g., a chain of events in which each action automatically causes other actions) but with no planning/no clear goal-directed behavior

Levels 1-2 cannot be coded for SG elements; you may use the non-narrative elements listed above (b) if you find it helpful (I will not be comparing these for reliability, however. Level 3 may have the SG elements of Settings, IE/Ps and Consequences)

At each of the following SG levels, the aims or goals of a character are either described or implied

iv. **Level 4**: Abbreviated episode – Story is goal-directed, but characters' intent is not explicitly stated and must be inferred; the main character may have a goal but make no plan or perform any intentional action (e.g., attempt) to attain it. For example, a story may present the problem of older children being bullies at school. Time passes, the younger, bullied children grow older, and the problem resolves. The answer to the question, "is planning or intentional behavior explicit?" for this story is no. Initiating events lead to consequences and resolutions with no attempt, planning, or stated intention on the part of the main character.

v. **Level 5:** Incomplete episode – Characters' intent is explicitly stated (e.g., the episode must contain either an internal plan or an attempt) but one of the following episode components is missing: initiating event/problem, attempt, or consequence.

vi. **Level 6:** Complete episode – Includes aims and plans of a character; may reflect evidence of planning in the attempts of a character to reach the goal; minimally has an initiating event/problem, attempt, and a consequence

vii. **Level 7:** Multiple episode – Is a chain of reactive sequences or abbreviated episodes, or a combination of complete and incomplete episodes

viii. **Level 8:** Complex episode – Full episode is elaborated by including multiple plans, attempts, or consequences; includes an obstacle to obtaining the goal (e.g., a trick as in "trickster tales")

ix. **Level 9:** Embedded episodes – One episode embedded within another

x. **Level 10:** Interactive episodes – Use multiple perspectives to describe events; multiple characters and multiple goals mutually influence each other; may have a reaction or consequence for one character serving as an initiating event for another character

2. For stories at the level of *abbreviated episode and higher* – code story grammar elements. As you read through the narrative again parsing out elements, it may be necessary to reformulate the utterances into propositions that function as particular elements (e.g., setting, initiating event or problem, attempt, etc.). Assign elements as applicable. Some utterances will not serve as any element. Do not code them but feel free to leave a note as to how you interpret them (e.g., redundant information, extraneous information, etc.). Use the following rules as a guide for coding SG elements:

    a. **Setting (S):** Reference to time and place, usually including introduction of one or more characters; a character's habitual state may be noted and/or a habitual social context

    b. **Initiating Event or Problem (IE/P):** An event, sometimes called "complication," that sets the events of the story in motion, including a problem that requires a solution; it functions to make the protagonist want to achieve a goal or change of state; IEs could be: (a) a character's action or an event, (b) natural occurrences, and (c) internal events, including a character's internal perception of an external event. Setting and IEs are distinguished from each other in that the Setting provides the context for

the story and the IE always evokes an immediate response from the character.

c. **Internal Response (IR):** The psychological state of the character after the initiating event. There are three types of potential responses: (a) affective response (emotional); (b) goals (references to a character's intended behavior); and (c) cognitions (statements that refer to a character's thoughts (e.g., "he remembered", "he thought", "he realized")

d. **Internal Plan (IP):** Statements that specify a character's strategy for obtaining a goal. There can be two aspects to a plan: (a) cognitions (thoughts about the situation or possible obstacles to the main goal, hypothesized activity, or consequences of behavior; (b) sub-goals (secondary goals to achieve main goal - can include if-then concepts – if the character first does this, then something else could occur that would bring him/her closer to the goal)

e. **Attempt (A):** Some action or a series of actions *taken by the main character* that is meant to solve the problem or attain the goal; there may be several attempts without a statement of consequence before the end of a story. An Attempt represents a character's overt action toward resolving the situation or achieving a goal. There needs to be a direct causal link or enablement relation between the Attempt and either the IE or IR that usually precedes it, or a direct causal link or enablement relation between the Attempt and subsequent Consequence.

f. **Consequence (C):** The success or failure of the character in achieving goal. With simple stories, the consequences may be a direct result of an initiating event. There are 3 types of consequences: (a) natural occurrences

– changes in the physical environment, usually not caused by an animate being; (b) action – physical activities carried out by animate characters that are the goal attainment; (c) End States – final state of the environment or characters (e.g., “They were happy inside the cabin”, “the town was left in ruins”).

g. **Resolution or Reaction (Res/Rea):** The character’s feelings, thoughts or actions in response to the consequences of attaining or not attaining a goal. There are 3 types of resolutions: (a) affect – the character’s emotional state; (b) cognition – the character’s thoughts; and (c) action – actions that result from the consequence or emotional responses. The final state or situation triggered by the initiating event; it does not cause or lead to other actions or states.

h. **Ending (E):** A sentence or phrase that clearly states the story is over (e.g., “the end”) or wraps up the story (e.g., provides a moral of the story, etc.)

APPENDIX B2: NARRATIVE ASSESSMENT PROFILE MODIFIED CODING CRITERIA

Narrative Aspect	Coding Criteria
Topic Maintenance	2 = Almost all utterances on topic (A) 1 = Most on topic; off-topic associations, likes, dislikes, etc. (V) 0 = Most utterances are off-topic and/or topic is difficult to discern (I)
Event Sequencing	2 = All events in chronological order or acknowledged as out of order by conjunctions, etc. (A) 1 = Most events chronologically ordered (V) 0 = No chronological ordering (or most not in order) (I)
Informativeness: Police Officer	2 = All specific information necessary to understand experience is provided or implied; credit should be given for easily inferred information (A) 1 = Most specific information provided but omissions of a few important points (for example, beginning, middle, or end), - leaving the listener with some questions as to what happened 0 = Not enough information (or too much information) to make sense of what happened (I)
Informativeness: Teacher	2 = Provides elaboration and embellishment of most important points of story including at least 2 of 3 ingredients (evaluation includes inferencing) (A) 1 = Provides some elaboration of some important points including at least 1 of the 3 ingredients (V) 0 = Provides little to no elaboration (1-2 statements at best) OR provides too much elaboration of extraneous details, detracting from

	storyline (I)
Informativeness: Chef	2 = All three ingredients must be present and appropriate in proportion; they provide sufficient information, leaving no gaps and/or creating no "noise" that impedes understanding (A) 1 = Two ingredients are present without important gaps (V) 0 = One or no ingredients are present without important gaps (I) Ingredients are: description (orientation), action, and evaluation
Referencing	2 = All references are cohesive; pronouns are used appropriately and their antecedents are clear (A) 1 = Most pronouns used appropriately and antecedents are mostly clear (V) 0 = Severely impaired referencing (includes inappropriate pronoun use and/or unspecified or confusing antecedents) (I)
Conjunctive Cohesion	2 = Variety of conjunctions are used appropriately (may include "and", "then", "but", "so", "because", etc.) (A) 1 = Mostly Only ands, and thens (V) 0 = No conjunctions or conjunctions are used inappropriately, impeding understanding (I)
Fluency	2 = Fluent: Most utterances are fluent; the few that are dysfluent don't interfere with understanding. Fewer than 20% of utterances are dysfluent. 1 = A few dysfluencies: Some dysfluencies but still comprehensible (Between 20-30% of utterances are dysfluent).

	0 = Mostly dysfluent (30% or more of utterances are dysfluent).

APPENDIX B3: NARRATIVE SCORING SCHEME MODIFIED RUBRIC

Characteristic	5 = Proficient	4	3 = Emerging	2	1 = Minimal
Introduction	Setting elements include introduction of character and occur before the character starts acting. Introduces at least 1 key character while providing a time and a place. Must provide some detail about the setting · setting (e.g., reference to the time of the setting, daytime, bedtime, season). Setting elements are stated at appropriate place in story.	Setting elements include introduction of main character but subsequent characters are not adequately introduced and/or referenced.	Must mention a main character AND provide either time, location, or use a setting statement like "once upon a time." One day a boy was playing baseball. OR There was a boy playing baseball. OR A boy was playing baseball in the street. 2 setting elements, one of which is an important character.	A setting statement without main character OR mentions main character without a setting statement (may appear later in story). (One day) there was a circus. A boy wanted to see a circus. A boy goes to the circus. A boy was playing baseball.	Launches into story with no attempt to provide the setting or introduce character(s) (e.g., 'They were running. He threw his popcorn.')
Character Development	Main characters are introduced with some description or detail provided.	Attempt to develop main character using either mental state words	Both main and active supporting characters are mentioned. There is a hierarchy such that	No main character – a number of characters may be mentioned but there is no hierarchy.	Characters are unable to be determined (e.g., all pronouns).

Characteristic	5 = Proficient	4	3 = Emerging	2	1 = Minimal
	Main character(s) and all supporting character(s) are mentioned. Throughout the story it is clear child can discriminate between main and supporting characters (e.g., more description of, emphasis upon main character(s)). Child narrates in first person using character voice (e.g., 'You get out of my tree', said the owl). Elaboration of character(s) that doesn't leave major questions/gaps	and/or dedicating a number of utterances to that character. (Mental states of important character definitely deserve 4). The listeners know something about this character. Consider number and kind of utterances devoted to that character.	a main character is distinguishable from supporting characters.		Are no characters. Only mention of general/collective character (everyone).
Mental States	Mental states of main and supporting characters are expressed when necessary for plot development and advancement (reason for why they felt that way -	A variety of mental state words are used without explanation.	Use of the same mental state word (multiple times). A singular mention of a mental state with a reason (he was scared because the lion	A singular mention of mental state.	No use of mental state words to develop character(s).

Characteristic	5 = Proficient	4	3 = Emerging	2	1 = Minimal
	because, so, etc.). A variety of mental state words are used. Reason has to be clearly marked, not implied (because, therefore, so).		was chasing him).		
Referencing	No ambiguity (characters, not events). AND Use of clarifiers (this one, that one, the other one) to enhance comprehensibility.	No ambiguity (characters, not events).	Inconsistent use of referents/antecedents. Some appropriate and some inappropriate referencing but inappropriate referencing doesn't interfere with comprehension of the basic story; listener is not confused. Comprehensibility doesn't suffer although some are still inappropriate.	Inconsistent (some appropriate/some inappropriate) use of clarifiers such that comprehensibility is compromised (the listener is confused). Some are appropriate but comprehensibility suffers.	Excessive use of pronouns. <u>No verbal clarifiers used when needed</u> or inappropriate use of indefinite/definite articles (a/the) and other clarifiers (this, that, these, those, etc.).
Conflict Resolution	Clearly states all conflicts and resolutions critical to advancing the plot of the story.	All critical conflicts and resolutions are present but they	There is at least one discernible conflict and its resolution; they may be under developed but can be logically	Random resolution(s) stated with no mention of cause or conflict.	Appears to primarily be a descriptive sequence with no

Characteristic	5 = Proficient	4	3 = Emerging	2	1 = Minimal
	All *critical/pivotal* conflicts are adequately resolved – does not leave listener hanging.	may be underdeveloped (but can be logically inferred).	inferred.	OR Conflict mentioned without resolution. OR Many conflicts and resolutions critical to advancing the plot are not present.	discernible conflict. .
Cohesion	Events follow a logical order. Critical events are included while less emphasis is placed on minor events. Smooth transitions are provided between events.	Events follow a logical order (unless violation of order is intentional and made abundantly clear). Some Appropriate use of a variety of subordinating conjunctions (and, then, so, because,	Events follow a logical order. Excessive detail or emphasis provided on minor events leading the listener astray. Equal emphasis on all events because of a lack of or inappropriate use of subordinating conjunctions and/or conjunctions are not used appropriately.	Within the sequence there is an attempt to connect utterances with active use of cohesive devices.	A series of disconnected utterances.

Characteristic	5 = Proficient	4	3 = Emerging	2	1 = Minimal
		therefore).			
Conclusion	Story is clearly wrapped up using general concluding statements such as 'and they were together again happy as could be'.	Story is clearly wrapped up using general concluding statements BUT a significant event or impact on a character is unresolved.	Specific event is concluded, but no general statement made as to the conclusion of the whole story.	If narrative is a descriptive sequence there is no event to conclude. If the child indicates that descriptive sequence is finished using a device such as "and that's it" or "that is all", give it a 2 (because s/he didn't just stop).	Stops narrating.

APPENDIX C: TRANSCRIPT CODING DECISION RULES

Decisions to be made:

1. Where to break utterances (use line break – one utterance per line)
2. What kind of end punctuation to give each utterance
3. Which words/phrases are mazes (use parentheses to separate mazes)
4. How many clauses are in an utterance (use [SI-__] code at end of utterance to indicate # of clauses) or whether an utterance cannot be coded for # of clauses due to unintelligible speech, etc. (in this case, use [SI-X]).
5. Whether or not a clause has any errors (could be word errors, word order errors, morpheme errors, etc.) (use [EU] before end punctuation to indicate error within the utterance).

Decision Rules:

1. Utterance break decisions:
    a. Generally, look for subject – predicate constructions and break after each one, unless one is subordinate to another.
    b. In the absence of subject – predicate constructions (e.g., fragments), rely on semantic or context cues to judge whether or not the fragment is associated with (and should be included with) the utterance either before or after it or whether it was an attempted utterance that was abandoned. If it was attempted yet abandoned, it receives its own line and will be punctuated with the abandoned utterance symbol, “>”.
2. End punctuation decisions:
    a. Is the utterance a complete sentence? -→ “.”
        i. Use a period if the utterance is a complete sentence (C-unit) even though it may have errors (e.g., “the boy run away from they lion.”

Even though the words “run” and “they” are errors, the utterance is a complete sentence with subject, verb and its argument).

b. Is the utterance incomplete? -→ “>”

   i. Use a greater than symbol if the utterance is incomplete/abandoned (for example, “then the boy was> the lion wanted to eat the boy.” The child initiated a sentence, “the boy was,” but never completed it. Put a “>” symbol after “was” to indicate that the utterance was abandoned and begin the next utterance on the subsequent line).

c. Was the child’s speech interrupted by the examiner: -→ “^”

   i. Use a carrot if the child’s interrupted speech flow is due to the examiner saying something. The next line would then reflect what the examiner said that interrupted the child (for example, “C He was^” / “E go on.” / “C He was running”).

d. Use “?” or “!” if those are in the original transcript – usually these occur in the context of a character’s speech enclosed in direct quotes within the narrative or in the context of dialogue between the child and the examiner.

3. <u>Maze decisions</u>:

   a. Are there words that are extraneous and unnecessary to the meaning of the utterance? -→ offset them in ( ).

      i. Mazes typically include false starts, reformulations, meaningless repetitions, etc. and are offset in parentheses so as not to inflate word count. Some things to consider:

         1. ***Are there repetitions that are intended to provide emphasis*** (for example, “the boy ran ran ran away from the lion”)? If so, give the child credit for two instances of the

word only. Connect the 2nd and any subsequent instances of the word with an underscore (e.g., "the boy ran ran_ran away from the lion").

2. ***Give credit for coordinating conjunctions even when they are used repeatedly throughout the story*** (for example, when a child begins every utterance with "and" or "then" or "and then" do not count those as mazes, *unless they are repeated within* an utterance (for example, "(and then) and then the boy went home").
3. ***When "like" or "you know" are used as fillers, they are maze***s. Usually, if they are behaving as fillers, they will appear repeatedly in the transcript (for example, "it was (like, you know, like) the day of a circus."). If it appears that these expressions carry semantic meaning or purpose in the story, do not offset them in parentheses (for example, "the boy ran like the wind" or "The cop grabbed the boy's wrist. You know that must have hurt." In this case "you know" is not a filler, rather it is being used as an evaluative device and so it has semantic value and therefore should be counted toward the word count).

4. <u>Counting clauses for the subordination index [SI].</u>
    a. Does the utterance contain one complete independent clause? → [SI-1]. Example: "and they all went home [SI-1]."
    b. Does the utterance contain an independent plus a dependent clause? → [SI-2] or if two dependent clauses, [SI-3], etc. Example: "when the lion

escaped, they all went home [SI-2].” Or, “when the lion escaped, they all went home because they were scared [SI-3].”

c. Do any of the clauses have sections that are unintelligible (marked with XX)? → Do not give credit for clauses with any unintelligible segments. Example: “when XX lion escaped, they all went home [SI-1].” – *give credit for the main (intelligible) clause but not the unintelligible one.* If both clauses have unintelligible segments, SI cannot be determined, therefore [SI-X].

d. Was the utterance abandoned? → SI cannot be determined, therefore [SI-X].

e. Conversational insertions/colloquialisms are also not scored for SI, therefore they receive [SI-X].

f. Is the clause missing an obligatory subject or verb? → [SI-0].

5. <u>Determining utterance error:</u>

a. Is the utterance a complete sentence and grammatically acceptable by native-English standards? → does not receive an [EU] code. Example: “(the boy um he) the boy went home [SI-1].”

b. Does the utterance have any errors that make it unacceptable by native-English speakers’ standards? → [EU]. Example: “(the boy um he) the boy go home [SI-1] [EU].” The verb, go, has agreement and/or tense error, therefore the utterance receives an error code.

c. Is there a main clause plus 1 or more subordinate clauses? → Indicate how many of the clauses have an error with a number after the [EU]. Example: “the boy go home because he was scared [SI-2] [EU1].”

d. When there are unintelligible parts of an utterance, error cannot be determined → [EUX].

References

Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research, 74*(1), 1-28.

Adlof, S. M., Catts, H. W., & Lee, J. (2010). Kindergarten predictors of second versus eighth grade reading comprehension impairments. *Journal of Learning Disabilities, 43*(4), 332-345.

Allen, M. S., Kertoy, M. K., Sherblom, J. C., & Pettit, J. M. (1994). Children's narrative productions: A comparison of personal event and fictional stories. *Applied Psycholinguistics, 15*(2), 149-176.

Applebee, A. (1978). *The child's concept of story: Ages two to seven*. Chicago, IL: University of Chicago Press.

Artiles, A., & Ortiz, A. A. (Eds.). (2002). *English Language Learners with Special Education Needs: Identification, Assessment, and Instruction*. McHenry, IL: CAL and Delta Systems Co., Inc.

Artiles, A. J. (2009). Re-framing disproportionality research: Outline of a cultural-historical paradigm. *Multiple Voices for Ethnically Diverse Exceptional Learners, 11*(2), 24-37.

Artiles, A. J. (2011). Toward an interdisciplinary understanding of educational equity and difference: The case of the racialization of ability. *Educational Researcher, 40*(9), 431-445.

Artiles, A. J., & Klingner, J. K. (2006). Forging a knowledge base on English language learners with special needs: Theoretical, population, and technical issues. *Teachers College Record, 108*(11), 2187-2194.

Artiles, A. J., Kozleski, E. B., Trent, S. C., Osher, D., & Ortiz, A. (2010). Justifying and explaining disproportionality, 1968-2008: A critique of underlying views of culture. *Exceptional Children, 76*(3), 279-299.

Artiles, A. J., Rueda, R., Salazar, J. J., & Higareda, I. (2005). Within-group diversity in minority disproportionate representation: English language learners in urban school districts. *Exceptional Children, 71*(3), 283-300.

Artiles, A. J., & Trent, S. C. (1994). Overrepresentation of minority students in special education: A continuing debate. *The Journal of Special Education, 27*(4), 410-437.

Aud, S., Hussar, W., Kena, G., Bianco, K., Frohlich, L., Kemp, J., & Tahan, K. (2011). *The condition of education 2011 (NCES 2011-033).* Washington, DC: U.S.: Government Printing Office.

August, D., & Hakuta, K. (1997). *Improving schooling for language minority children: A research agenda*. Washington, DC: National Research Council and Institute of Medicine, National Academy Press.

August, D., & Shanahan, T. (Eds.). (2006). *Developing reading and writing in second-language learners: Report of the National Literacy Panel on language-minority children and youth*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bedore, L. M., & Peña, E. D. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education & Bilingualism, 11*(1), 1-29.

Bedore, L. M., Peña, E. D., Gillam, R. B., & Ho, T. (2010). Language sample measures and language ability in Spanish-English bilingual kindergarteners. *Journal of Communication Disorders, 43*(6), 498-510.

Berman, R. A., & Slobin, D. I. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc.

Bialystok, E. (2007). Acquisition of literacy in bilingual children: A framework for research. *Language Learning, 57*(1), 45-77.

Blanchett, W. J., Mumford, V., & Beachum, F. (2005). Urban school failure and disproportionality in a post-*Brown* era: Benign neglect of the constitutional rights of students of color. *Remedial & Special Education, 26*(2), 70-81.

Bliss, L. S., McCabe, A., & Miranda, A. E. (1998). Narrative Assessment Profile: Discourse analysis for school-age children. *Journal of Communication Disorders, 31*, 347-363.

Botvin, G. J., & Sutton-Smith, B. (1977). The development of structural complexity in children's fantasy narratives. *Developmental Psychology, 13*(4), 377-388.

Boudreau, D. (2008). Narrative abilities: Advances in research and implications for clinical practice. *Topics in Language Disorders, 28*(2), 99-114.

Boudreau, D. M., & Hedberg, N. L. (1999). A comparison of early literacy skills in children with specific language impairment and their typically developing peers. *American Journal of Speech-Language Pathology, 8*, 249-260.

Celinska, D. K. (2004). Personal narratives of students with and without learning disabilities. *Learning Disabilities Research & Practice, 19*(2), 83-98.

Cleave, P. L., Girolametto, L. E., Chen, X., & Johnson, C. J. (2010). Narrative abilities in monolingual and dual language learning children with specific language impairment. *Journal of Communication Disorders, 43*(6), 511-522.

Cole, M. (1996). *Cultural psychology: A once and future discipline*. Cambridge, MA: Belknap Press of Harvard University Press.

Coutinho, M. J., & Oswald, D. P. (2000). Disproportionate representation in special education: A synthesis and recommendations. *Journal of Child and Family Studies, 9*(2), 135-156.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*(2), 222-251.

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy*. Clevedon, Avon, England: Multilingual Matters.

Cummins, J. (2001). Empowering minority students: A framework for intervention. *Harvard Educational Review, 71*(4), 649-675.

Damico, J. S. (1991). Descriptive assessment of communicative ability in limited English proficient students. In E. V. Hamayan & J. S. Damico (Eds.), *Limiting bias in the assessment of bilingual students* (pp. 157-218). Austin: PRO-ED.

Damico, J. S., Oller, J. W., & Storey, M. E. (1983). The diagnosis of language disorders in bilingual children: Surface-oriented and pragmatic criteria. *Journal of Speech and Hearing Disorders, 48*, 385-394.

De Avila, E. A., & Duncan, S. E. (1990). *Language Assessment Scales Oral - English*. Monterey, CA: CTB/McGraw Hill.

De Valenzuela, J. S., Copeland, S. R., Huaqing Qi, C., & Park, M. (2006). Examining educational equity: Revisiting the disproportionate representation of minority students in special education. *Exceptional Children, 72*(4), 425-441.

Delpit, L. (2006). *Other people's children: Cultural conflict in the classroom*. New York: The New Press.

Deno, E. (1970). Special education as developmental capital. *Exceptional Children, 37*, 229-237.

Dickinson, D. K., & McCabe, A. (2001). Bringing it all together: The multiple origins, skills, and environmental supports of early literacy. *Learning Disabilities Research & Practice, 16*(4), 186-202.

Donovan, S., & Cross, C. (2002). *Minority students in special and gifted education*: National Academies Press.

Dunn, L. M. (1968). Special education for the mildly retarded: Is much of it justifiable? *Exceptional Children, 23*, 5-21.

Eisner, E. W. (2003). Questionable assumptions about schooling. *Phi Delta Kappan, 84*(9), 648-657.

Ely, R., Wolf, A., McCabe, A., & Melzi, G. (2000). The story behind the story: Gathering narrative data from children. In L. Menn & N. B. Ratner (Eds.), *Methods for studying language production*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ennis, S. R., Ríos-Vargas, M., & Albert, N. G. (2012). *La población Hispana: 2010*. Washington, DC.

Evans, R. (2005). Reframing the achievement gap. *Phi Delta Kappan*, 582-589.

Fazio, B. B., & Naremore, R. C. (1996). Tracking children from poverty at risk for specific language impairment: a 3-year longitudinal study. *Journal of Speech and Hearing Research, 39*(3), 611-625.

Fiestas, C. E., & Peña, E. D. (2004). Narrative discourse in bilingual children: Language and task effects. *Language, Speech, and Hearing Services in Schools, 35*(2), 155-168.

Figueroa, R. A. (2002). Toward a new model of assessment. In A. Artiles & A. A. Ortiz (Eds.), *English language learners with special education needs: Identification, assessment, and instruction* (pp. 51-64). Washington, DC: Center for Applied Linguistics and Delta Systems.

Fitzgerald, J. (1984). The relationship between reading ability and expectations for story structures. *Discourse Processes, 7*, 21-41.

Francis, D. J., Carlson, C. D., Fletcher, J. M., Foorman, B. R., Goldenberg, C. R., & Vaughn, S. (2005). Oracy / Literacy development of Spanish-speaking children: A multi-level program of research on language minority children and the instruction, school and community contexts, and interventions that influence their academic outcomes. *The International Dyslexia Association, 31*, 8-12.

Garcia, E., Arias, M. B., Murri, N. J. H., & Serna, C. (2010). Developing responsive teachers: A challenge for a demographic reality. *Journal of Teacher Education, 61*((1-2)), 132-142.

Garcia, E., & Cuéllar, D. (2006). Who are these linguistically and culturally diverse students? *Teachers College Record, 108*(11), 2220-2246.

Garcia, S. B. (2002). Parent-professional collaboration in culturally sensitive assessment. In A. Artiles & A. A. Ortiz (Eds.), *English language learners with special education needs: Identification, assessment, and instruction* (pp. 87-103). Washington, DC: Center for Applied Linguistics and Delta Systems.

García, S. B., & Guerra, P. L. (2004). Deconstructing deficit thinking: Working with educators to create more equitable learning environments. *Education and Urban Society, 36*(2), 150-168.

Garcia, S. B., & Ortiz, A. A. (2006). Preventing disproportionate representation: Culturally and linguistically responsive prereferral interventions. *Teaching Exceptional Children, 38*(4), 64-68.

Gazella, J., & Stockman, I. (2003). Children's story retelling under different modality and task conditions: Implications for standardizing language sampling procedures. *American Journal of Speech-Language Pathology, 12*(1), 61-73.

Glenn, C., & Stein, N. (1980). *Syntactic structures and real world themes in stories generated by children*. Urbana, IL: University of Illinois Center for the Study of Reading.

Goldstein, B. C., Harris, K. C., & Klein, M. D. (1993). Assessment of oral storytelling abilities of Latino junior high school students with learning handicaps. *Journal of Learning Disabilities, 26*(2), 138-143.

Gorman, B. K., Fiestas, C. E., Peña, E. D., & Clark, M. R. (2011). Creative and stylistic devices employed by children during a storybook narrative task: A cross-cultural study. *Language, Speech, and Hearing Services in Schools, 42*, 167-181.

Greenfield, P. M., Keller, H., Fuligni, A., & Maynard, A. (2003). Cultural pathways through universal development. *Annual Review of Psychology, 54*, 461-490. doi: 10.1146/annurev.psych.54.101601.145221

Griffin, T. M., Hemphill, L., Camp, L., & Wolf, D. P. (2004). Oral discourse in the preschool years and later literacy skills. *First Language, 24*(71), 123-147.

Gudyknust, W. B., Matsumoto, Y., Ting-Toomey, S., Nishida, T., Kim, K., & Heyman, S. (1996). The influence of cultural individualism-collectivism, self construals, and individual values on communication styles across cultures. *Human Communication Research, 22*(4), 510-543.

Gutiérrez-Clellen, V. F. (1995). Narrative development and disorders in Spanish-speaking children: Implications for the bilingual interventionist. In H. Kayser (Ed.), *Bilingual Speech-Language Pathology*. San Diego: Singular Publishing Group, Inc.

Gutiérrez-Clellen, V. (1998). Syntactic skills of Spanish-speaking children with low school achievement. *Language, Speech, & Hearing Services in Schools, 29*, 207-215.

Gutiérrez-Clellen, V. F. (2002). Narratives in two languages: Assessing performance of bilingual children. *Linguistics and Education, 13*(2), 175-197.

Gutiérrez-Clellen, V., & Hofstetter, R. (1994). Syntactic complexity in Spanish narratives: A developmental study. *Journal of Speech and Hearing Research, 37*, 645-654.

Gutiérrez-Clellen, V., & Iglesias, A. (1992). Causal coherence in the oral narratives of Spanish-speaking children. *Journal of Speech and Hearing Research, 35*, 363-372.

Gutiérrez-Clellen, V. F., Simon-Cereijido, G., & Wagner, C. (2008). Bilingual children with language impairment: A comparison with monolinguals and second language learners. *Applied Psycholinguistics, 29*(1), 3-19.

Gwet, K. L. (2011). On the Krippendorff's alpha coefficient. Retrieved from http://www.agreestat.com/research_papers/onkrippendorffalpha.pdf

Harry, B., & Klingner, J. (2006). *Why are so many minority students in special education? Understanding race & disability in schools.* New York: Teachers College Press.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1,* 77-89.

Haynes, W. O., Haynes, M. D., & Strickland-Helms, D. F. (1989). Alpha hemispheric asymmetry in children with learning disabilities and normally achieving children during story comprehension and rehearsal prior to narrative production. *Journal of Learning Disabilities, 22*(6), 391-399.

Hayward, D. V., Gillam, R. B., & Lien, P. (2007). Retelling a script-based story: Do children with and without language impairments focus on script and story elements? *American Journal of Speech-Language Pathology, 16*, 235-245.

Hedberg, N., & Westby, C. (1993). *Analyzing storytelling skills: Theory to practice*. Tucson, AZ: Communication Skill Builders.

Heilmann, J., Miller, J. F., & Nockerts, A. (2010). Sensitivity of narrative organization measures using narrative retells produced by young school-age children. *Language Testing, 27*(4), 603-626.

Heilmann, J., Miller, J. F., Nockerts, A., & Dunaway, C. (2010). Properties of the Narrative Scoring Scheme using narrative retells in young school-age children. *American Journal of Speech and Language Pathology, 19,* 154-166.

Hemphill, L. (1989). Topic development, syntax, and social class. *Discourse Processes, 12,* 267-286.

Hudson, J. A., & Shapiro, L. R. (1991). From knowing to telling: The development of children's scripts, stories, and personal narratives. In A. McCabe & C. Peterson (Eds.), *Developing narrative structure* (pp. 89-136). Hillsdale, N.J.: Erlbaum Associates.

Hughes, D., McGillivray, L., & Schmidek, M. (1997). *Guide to narrative language: Procedures for assessment*. Austin, TX: PRO-ED.

IBM Corp. (2011). IBM SPSS Statistics for Mac (Version 21.0). Armonk, NY: IBM Corp.

Kaderavek, J. N., & Sulzby, E. (2000). Narrative production by children with and without specific language impairment: oral narratives and emergent readings. [Feature Article]. *Journal of Speech, Language, and Hearing Research, 43*(1), 34-49.

Kindler, A. L. (2002). *Survey of the states' limited English proficient students and available educational programs and services, 2000-2001 summary report*. Washington, DC: National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs.

Klecan-Aker, J. S., & Kelty, K. R. (1990). An investigation of the oral narratives of normal and language-learning disabled children. *Communication Disorders Quarterly, 13*(2), 207-216.

Klingner, J., Hoover, J. J., & Baca, L. M. (Eds.). (2008). *Why do English language learners struggle with reading? Distinguishing language acquisition from learning disabilities*. Thousand Oaks, CA: Corwin Press.

Klingner, J. K., Artiles, A. J., & Barletta, L. M. (2006). English Language Learners Who Struggle with Reading: Language Acquisition or LD? *Journal of Learning Disabilities, 39*(2), 108-128.

Klingner, J. K., Artiles, A. J., Kozleski, E., Harry, B., Zion, S., Tate, W., . . . Riley, D. (2005). Addressing the disproportionate representation of culturally and linguistically diverse students in special education through culturally responsive educational systems. *Education Policy Analysis Archives, 13*(38), 1-43.

Klingner, J. K., & Harry, B. (2006). The special education referral and decision-making process for English language learners: Child study team meetings and placement conferences. *Teachers College Record, 108*(11), 2247-2281.

Krezmien, M. P., Leone, P. E., & Achilles, G. M. (2006). Suspension, race, and disability: Analysis of statewide practices and reporting. *Journal of Emotional & Behavioral Disorders, 14*(4), 217-226.

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Labov, W. (1972). *Language in the Inner City*. Philadelphia: University of Pennsylvania Press.

Laing, S. P., & Kamhi, A. (2003). Alternative assessment of language and literacy in culturally and linguistically diverse populations. *Language, Speech, & Hearing Services in Schools, 34*, 44-55.

Leone, P. E., Christle, C. A., Nelson, C. M., Skiba, R., Frey, A., & Jolivette, K. (2003). School failure, race, and disability: Promoting positive outcomes, decreasing vulnerability for involvement with the juvenile delinquency system. Washington, D.C.: EDJJ: The National Center on Education, Disability, and Juvenile Justice.

Liles, B. Z., Duffy, R. J., Merritt, D. D., & Purcell, S. L. (1995). Measurement of narrative discourse ability in children with language disorders. *Journal of Speech and Hearing Research, 38*, 415-425.

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

Losen, D. J., & Welner, K. G. (2001). Disabling discrimination in our public schools: Comprehensive legal challenges to inappropriate and inadequate special education

services for minority children. *Harvard Civil Rights-Civil Liberties Law Review, 36*, 407-460.

MacMillan, D. L., & Reschly, D. J. (1998). Overrepresentation of minority students: The case for greater specificity or reconsideration of the variables examined. *Journal of Special Education, 32*(1), 15-24.

Mandler, J. M. (1988). A code in the node: The use of a story schema in retrieval. In M. B. Franklin & S. S. Barten (Eds.), *Child language: A reader* (pp. 278-281). New York: Oxford University Press.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology, 9*, 111-151.

Manhardt, J., & Rescorla, L. (2002). Oral narrative skills of late talkers at ages 8 and 9. *Applied Psycholinguistics, 23*(1), 1-21.

Martinez-Roldán, C. M., & Sayer, P. (2006). Reading through linguistic borderlands: Latino students' transactions with narrative texts. *Journal of Early Childhood Literacy, 6*(3), 293-322.

Marx, S. (2004). Regarding whiteness: Exploring and intervening in the effects of white racism in teacher education. *Equity & Excellence in Education, 37*, 31-43.

McCabe, A., Bailey, A. L., & Melzi, G. (Eds.). (2008). *Spanish-language narration and literacy*. Cambridge, MA: Cambridge University Press.

McCabe, A., & Bliss, L. S. (2003). *Patterns of narrative discourse: A multicultural, life span approach*. Boston, MA: Pearson.

McCabe, A., & Rollins, P. R. (1994). Assessment of preschool narrative skills. *American Journal of Speech and Language Pathology, 3*, 45-56.

McCardle, P., Mele-McCarthy, J., & Leos, K. (2005). English language learners and learning disabilities: Research agenda and implications for practice. *Learning Disabilities Research & Practice, 20*(1), 68-78.

McCord, J. S., & Haynes, W. O. (1988). Discourse errors in students with learning disabilities and their normally achieving peers: Molar versus molecular views. *Journal of Learning Disabilities, 21*(4), 237-243.

McFadden, T. U., & Gillam, R. B. (1996). An examination of the quality of narratives produced by children with language disorders. *Language, Speech, & Hearing Services in Schools, 27*, 48-56.

McFarland, L. A. (2011). *Tell me a story: Narratives of Spanish-speaking English language learners*. Research Synthesis. Department of Special Education. The University of Texas at Austin.

McKown, C., & Weinstein, R. S. (2008). Teacher expectations, classroom context, and the achievement gap. *Journal of School Psychology, 46*(3), 235-261.

Merritt, D., & Liles, B. (1987). Story grammar ability in children with and without language disorder: Story generation, story retelling, and story comprehension. *Journal of Speech, Language, and Hearing Research, 30*(4), 539-552.

Miller, J. F., Andriacchi, K., & Nockerts, A. (2011). *Assessing language production using SALT software*. Middleton, WI: SALT Software, LLC.

Miller, J., & Iglesias, A. (2008). *Systematic Analysis of Language Transcripts (SALT) research version 2008*. Madison, WI: SALT Software LLC.

Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., & Francis, D. J. (2006). Oral language and reading in bilingual children. *Learning Disabilities Research & Practice, 21*(1), 30-43.

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

Montague, M., Maddux, C. D., & Dereshiwsky, M. I. (1990). Story grammar and comprehension and production of narrative prose by students with learning disabilities. *Journal of Learning Disabilities, 23*(3), 190-198.

Montanari, S. (2004). The development of narrative competence in the L1 and L2 of Spanish-English bilingual children. *International Journal of Bilingualism, 8*(4), 449-497.

Morris-Friehe, M., & Sanger, D. (1992). Language samples using three story elicitation tasks and maturation effects. *Journal of Communication Disorders, 25*(2), 107-124.

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, & Hearing Services in Schools, 34*(4), 332-342.

Nasir, N. S., & Hand, V. M. (2006). Exploring sociocultural perspectives on race, culture and learning. *Review of Educational Research, 76*(4), 449-475.

National Clearinghouse for English Language Acquisition (NCELA). (2011). What languages do English learners speak? *NCELA Fact Sheet*. Washington, DC: Author.

Ogbu, J. U. (1992). Understanding cultural diversity and learning. *Educational Researcher, 21*(8), 5-14+24.

Ogbu, J. U., & Simons, H. D. (1998). Voluntary and involuntary minorities: A cultural-ecological theory of school performance with some implications for education. *Anthropology & Education Quarterly, 29*(2), 155-188.

Ortiz, A. A. (1997). Learning disabilities occurring concomitantly with linguistic differences. *Journal of Learning Disabilities, 30*(3), 321-332.

Ortiz, A. A., Robertson, P. M., Wilkinson, C. Y., Liu, Y.-J., McGhee, B. D., & Kushner, M. I. (2011). The role of bilingual education teachers in preventing inappropriate referrals of ELLs to special education: Implications for Response to Intervention. *Bilingual Research Journal: The Journal of the National Association for Bilingual Education, 34*(3), 316-333.

Ortiz, A. A., Wilkinson, C. Y., Robertson-Courtney, P., & Kushner, M. I. (2006). Considerations in implementing intervention assistance teams to support English language learners. *Remedial & Special Education, 27*(1), 53-63.

Ortiz, A. A., & Yates, J. R. (2001). A framework for serving English language learners with disabilities. *Journal of Special Education Leadership, 14*(2), 72-80.

Ortiz, A. A., & Yates, J. R. (2002). Considerations in the assessment of English language learners referred to special education. In A. Artiles & A. A. Ortiz (Eds.), *English language learners with special education needs: Identification, assessment, and instruction* (pp. 65-86). Washington, DC: Center for Applied Linguistics and Delta Systems.

Owens, R. E. (2010). *Language disorders: A functional approach to assessment and intervention* (5th ed.). Boston: Pearson/Allyn & Bacon.

Patton, J. M. (1998). The disproportionate representation of African Americans in special education: Looking behind the curtain for understanding and solutions. *The Journal of Special Education, 32*(1), 25-31.

Paul, R., & Smith, R. L. (1993). Narrative skills in 4-year-olds with normal, impaired, and late-developing language. *Journal of Speech and Hearing Research, 36*, 592-598.

Pearce, W. (2003). Does the choice of stimulus affect the complexity of children's oral narratives? *International Journal of Speech-Language Pathology, 5*(2), 95-103.

Pearce, W. M., McCormack, P. F., & James, D. G. H. (2003). Exploring the boundaries of SLI: Findings from morphosyntactic and story grammar analyses. *Clinical Linguistics & Phonetics, 17*(4-5), 325-334.

Peterson, C., & McCabe, A. (1983). *Developmental psycholinguistics: Three ways of looking at a child's narrative*. New York: Plenum Press.

Peterson, D., Gillam, S., & Gillam, R. (2008). Emerging procedures in narrative assessment: The index of narrative complexity. *Topics in Language Disorders, 28*(2), 115-130.

Pray, L. (2005). How well do commonly used language instruments measure English oral-language proficiency? *Bilingual Research Journal, 29*(2), 387-409.

Reilly, J., Losh, M., Bellugi, U., & Wulfeck, B. (2004). "Frog, where are you?" Narratives in children with specific language impairment, early focal brain injury, and Williams syndrome. *Brain and Language, 88*, 229-247.

Ripich, D. N., & Griffith, P. L. (1988). Narrative abilities of children with learning disabilities and nondisabled children: Story structure, cohesion, and propositions. *Journal of Learning Disabilities, 21*(3), 165-173.

Rogoff, B. (2003). *The cultural nature of human development*. New York, NY: Oxford University Press.

Rogoff, B., & Chavajay, P. (1995). What's become of research on the cultural basis of cognitive development? *American Psychologist, 50*(10), 859-877.

Roth, F. P., Speece, D. L., & Cooper, D. H. (2002). A longitudinal analysis of the connection between oral language and early reading. *The Journal of Educational Research, 95*(5), 259-272.

Roth, F. P., Speece, D. L., Cooper, D. H., & De La Paz, S. (1996). Unresolved mysteries: How do metalinguistic and narrative skills connect with early reading? *Journal of Special Education, 30*(3), 257.

Roth, F. P., & Spekman, N. J. (1986). Narrative discourse: Spontaneously generated stories of learning-disabled and normally achieving students. *Journal of Speech and Hearing Disorders, 51*, 8-23.

Roth, F. P., Spekman, N. J., & Fye, E. C. (1995). Reference cohesion in the oral narratives of students with learning disabilities and normally achieving students. *Learning Disability Quarterly, 18*(1), 25-40.

Rumberger, R. W., & Larson, K. A. (1998). Toward explaining differences in educational achievement among Mexican American language-minority students. *Sociology of Education, 71*(1), 68-92.

Schneider, P. (1996). Effects of pictures versus orally presented stories on story retellings by children with language impairment. *American Journal of Speech-Language Pathology, 5*, 86-96.

Schneider, P., & Dubé, R. (2005). Story presentation effects on children's retell content. *American Journal of Speech-Language Pathology*, *14*(52-60).

Schoenbrodt, L., Kerins, M., & Gesell, J. (2003). Using narrative language intervention as a tool to increase communicative competence in Spanish-speaking children. *Language, Culture and Curriculum, 16*(1), 48-59.

Seidl, B., & Pugach, M. (2009). Support and teaching in the vulnerable moments: Preparing special educators for diversity. *Multiple Voices for Ethnically Diverse Exceptional Learners, 11*(2), 57-75.

Shiro, M. (2003). Genre and evaluation in narrative development. *Journal of Child Language, 30*(1), 165-195.

Silvaroli, N. J., Skinner, J. T., & Maynes, J. O. (1977). *Oral language evaluation*. St. Paul, MN: EMC Corporation.

Simon-Cereijido, G., & Gutiérrez-Clellen, V. F. (2009). A cross-linguistic and bilingual evaluation of the interdependence between lexical and grammatical domains. *Applied Psycholinguistics, 30*, 315-337.

Skiba, R. J., Michael, R. S., Nardo, A. C., & Peterson, R. (2000). The color of discipline: Sources of racial and gender disproportionality in school punishment: Indiana Education Policy Center.

Skiba, R. J., Poloni-Staudinger, L., Simmons, A. B., Feggins-Azziz, L. R., & Chung, C.-G. (2005). Unproven links: Can poverty explain ethnic disproportionality in special education? *The Journal of Special Education, 39*(3), 130--144.

Skiba, R. J., Simmons, A. B., Ritter, S., Gibb, A. C., Rausch, M. K., Cuadrado, J., & Choong-Geun, C. (2008). Achieving equity in special education: History, status, and current challenges. *Exceptional Children, 74*(3), 264-288.

Snyder, L. S., & Downey, D. M. (1991). The language-reading relationship in normal and reading-disabled children. *Journal of Speech and Hearing Research, 34*, 129-140.

Snyder, T. D., & Dillow, S. A. (2011a). *Digest of Education Statistics 2010 (NCES 2011-015).* National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Snyder, T. D., & Dillow, S. A. (2011b). Table 21. Estimates of resident population, by race/ethnicity and age group: Selected years, 1980 through 2010. *Digest of Education Statistics: 2010* Retrieved May 25, 2012, from http://nces.ed.gov/programs/digest/d10/tables/dt10_021.asp?referrer=list

Snyder, T. D., & Dillow, S. A. (2011c). Table 43. Percentage distribution of enrollment in public elementary and secondary schools, by race/ethnicity and state or jurisdiction: Fall 1998 and fall 2008. *Digest of Education Statistics: 2010* Retrieved May 25, 2012, from http://nces.ed.gov/programs/digest/d10/tables/dt10_043.asp?referrer=list

Snyder, T. D., & Dillow, S. A. (2011d). Table 115. Percentage of high school dropouts among persons 16-24 years old (status dropout rate), by sex and race/ethnicity: Selected years, 1960 through 2009. *Digest of Education Statistics: 2010* Retrieved May 25, 2012, from http://nces.ed.gov/programs/digest/d10/tables/dt10_115.asp?referrer=list

Snyder, T. D., & Dillow, S. A. (2011e). Table 132. Average reading scale scores of 4th- and 8th-graders in public schools and percentage scoring at or above selected reading achievement levels, by English language learner (ELL) status and state or jurisdiction: 2009. *Digest of Education Statistics: 2010* Retrieved May 25, 2012, from http://nces.ed.gov/programs/digest/d10/tables/dt10_132.asp?referrer=list

Speece, D. L., Roth, F. P., Cooper, D. H., & De La Paz, S. (1999). The relevance of oral language skills to early literacy: A multivariate analysis. *Applied Psycholinguistics, 20*(2), 167-190.

Spinillo, A., & Pinto, G. (1994). Children's narratives under different conditions: A comparative study. *British Journal of Developmental Psychology, 12*, 177-193.

Stein, N. (1988). The development of storytelling skill. In M. Franklin & S. Barten (Eds.), *Child language: A reader*. New York: Oxford University Press.

Stein, N. L. (1988). The development of children's storytelling skill. In M. B. Franklin & S. S. Barten (Eds.), *Child language: A reader* (pp. 282-297). New York: Oxford University Press.

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), *New Directions in Discourse Processing* (Vol. II). Norwood, NJ: Ablex Publishing Corporation.

Stockman, I. J. (1996). The promises and pitfalls of language sample analysis as an assessment tool for linguistic minority children. *Language, Speech, & Hearing Services in Schools, 27*, 355-366.

Trabasso, T., Stein, N., & Johnson, L. R. (1981). Children's knowledge of events: A causal analysis of story structure. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 15, pp. 237-282). New York: Academic Press.

Trueba, H. T. (1988). Culturally based explanations of minority students' academic achievement. *Anthropology & Education Quarterly, 19*(3), 270-287.

U.S. Department of Commerce Census Bureau. (2009). American Community Survey Retrieved July 22, 2011, from http://nces.ed.gov/programs/coe/tables/table-lsm-2.asp

U.S. Department of Education, National Center for Education Statistics, & Common Core of Data (CCD). (2012). "Local Education Agency Universe Survey," 2002-03 through 2010-11. Retrieved March 1, 2013, from http://nces.ed.gov/programs/digest/d12/tables/dt12_00n.asp

Uccelli, P., & Paez, M. M. (2007). Narrative and vocabulary development of bilingual children From kindergarten to first grade: Developmental changes and associations among English and Spanish skills. [Feature Article]. *Language, Speech, and Hearing Services in Schools, 38*(3), 225-236.

Uchikoshi, Y. (2005). Narrative development in bilingual kindergarteners: Can Arthur help? *Developmental Psychology, 41*(3), 464-478.

United States Census Bureau. (2006). Hispanics in the United States. Retrieved May 21, 2012, from http://www.census.gov/population/www/socdemo/hispanic/hispanic_pop_presentation.html

Villegas, A. M., & Lucas, T. (2002). Preparing culturally responsive teachers: Rethinking the curriculum. *Journal of Teacher Education, 53*(1), 20-32.

Wagner, R. K., Francis, D. J., & Morris, R. D. (2005). Identifying English language learners with learning disabilities: Key challenges and possible approaches. *Learning Disabilities Research & Practice, 20*(1), 6-15.

Weinstein, R. S., Gregory, A., & Strambler, M. J. (2004). Intractable self-fulfilling prophecies fifty years after Brown v. Board of Education. *American Psychologist, 59*(6), 511-520.

Westby, C. (1992). Narrative analysis. In W. A. Secord & J. S. Damico (Eds.), *Best practices in school speech-language pathology: Descriptive/nonstandardized language assessment* (pp. 91-101). San Antonio: The Psychological Corporation.

Westby, C. E. (1984). Development of narrative language abilities. In G. P. Wallach & K. G. Butler (Eds.), *Language learning disabilities in school age children* (pp. 213). Baltimore, MD: Williams and Wilkins.

Wilkinson, C. Y., Ortiz, A. A., Robertson, P. M., & Kushner, M. I. (2006). English language learners With reading-related LD: Linking data from multiple sources to make eligibility determinations. *Journal of Learning Disabilities, 39*(2), 129-141.

Wolman, C., van den Broek, P., & Lorch, R. F. (1997). Effects of causal structure on immediate and delayed story recall by children with mild mental retardation, children with learning disabilities, and children without disabilities. *The Journal of Special Education, 30*(4), 439-455.

Zehler, A. M., Fleischman, H. L., Hopstock, P. J., Pendzick, M. L., & Stephenson, T. G. (2003). Special topic report #4: Findings on special education LEP students *Descriptive study of services to LEP students and LEP students with disabilities*. Washington, D.C.: U.S. Department of Education.