AN ALTERNATIVE ORAL PROFICIENCY AND EXPRESSIVE VOCABULARY ASSESSMENT OF KINDERGARTEN ENGLISH LANGUAGE LEARNERS

A Dissertation

by

MIRANDA FERNANDE WALICHOWSKI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2009

Major Subject: Educational Psychology

AN ALTERNATIVE ORAL PROFICIENCY AND EXPRESSIVE VOCABULARY ASSESSMENT OF KINDERGARTEN ENGLISH LANGUAGE LEARNERS

A Dissertation

by

MIRANDA FERNANDE WALICHOWSKI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Rafael Lara-Alecio
Committee Members,	Sharolyn Pollard-Durodola
	Fuhui Tong
	Patricia Goodson
	Beverly Irby
Head of Department,	Victor Willson

December 2009

Major Subject: Educational Psychology

ABSTRACT

An Alternative Oral Proficiency and Expressive Vocabulary Assessment of Kindergarten English Language Learners.

(December 2009)

Miranda Fernande Walichowski, B.S., M.Ed., Texas A&M University

Chair of Advisory Committee: Dr. Rafael Lara-Alecio

The data used in this study were secondary, kindergarten data from a longitudinal, five-year, federal experimental research project: English and Literacy Acquisition (ELLA) (R305P030032). The overall goal of ELLA was to examine the impacts of two different programs (Bilingual and Structured English Immersion) on the performance of Spanish-speaking English language learners (ELLs) in grades K to 3.

My first research question was to determine to what extent a curriculum-based measure could be developed and validated to measure oral proficiency and vocabulary knowledge among ELLs who are participating in a controlled oral language development intervention. In addressing validity the scores of the S4 were compared with the scores of the Woodcock Language Proficiency Battery – Revised (WLPB-R) and the IOWA Test of Basic Skills (ITBS) language and vocabulary subtests. The correlations were .283 to .445 and they were statistically significant ($p<.01$). The S4 underwent several iterations. With each

iteration intrarater reliability improved (Kappa .817 to 1.00 and Cramer's V .330 to 1.00). Interrater reliability also improved (Kappa .431 to 1.00 and Cramer's V .616 to 1.00).

The second research question was to determine to what extent teachers could use the Semantic and Syntactic Scoring System (S4) for the STELLA vocabulary fluency measure with minimal training to accurately assess students' vocabulary knowledge and oral proficiency. The teachers' Kappas ranged from .786 to 1.00 and Cramer's V from .822 to 1.00. On average they were able to score a given student measure in under 22 minutes.

The third research question was to determine to what extent the Semantic and Syntactic Scoring System (S4) differentiates the level of knowledge regarding expressive vocabulary and oral proficiency of kindergarten students under two different program placements: enhanced Traditional Bilingual Education and the enhanced Structured English Immersion Program in comparison to the WLPB-R (language and vocabulary subtests). The S4 was able to distinguish between the control and experimental groups (unlike the other subtests); but was not able to distinguish program type (bilingual and structured English immersion).

DEDICATION

I dedicate this dissertation first and foremost to God. As I cross the river that is my life, He has always provided stepping-stones along the way. As waters rise and become turbulent and murky, He always unveils the next stone by calming and receding the waters, and I know that I can take the next step in faith. Second, I dedicate this dissertation to my husband, John. This dissertation is as much his accomplishment as it is mine because of his support, encouragement, and willingness to take on more than his share of the quotidian aspects of our life to afford me the time that I needed to work on this. Also, I dedicate this dissertation to my children, Brener and Bourne, who are the impetus for my wanting to be a better person. These three individuals are my constant reminder of God's unconditional love.

ACKNOWLEDGEMENTS

I acknowledge my parents, Fernando and Carmen Nava, my sisters, Ofelia and Dora, and my aunt, Josefina, because they have always believed in me. I also acknowledge my family by marriage: the Walichowskis South, the Sokoloskis, and the Anastases, and Aunt Tommy. I also appreciate them for being understanding when I have had to drop out from life so that I could write. I thank my family for all their prayers and support.

I am grateful to the principle investigators of Project ELLA, Dr. Rafael Lara, Dr. Beverly Irby, and Dr. Patricia Mathes, who allowed me to use data from their federally-funded, longitudinal study, ELLA. I further acknowledge my committee chair, Dr. Lara-Alecio, and Dr. Irby, committee member. It was their encouragement and confidence in me that lead me to pursue this degree. They have always provided mentorship and wisdom about life in academia. Dr. Goodson has influenced my life through the lessons that I learned from her in regards to creating a writing life. I acknowledge Dr. Durodola for providing her expertise and guidance in oracy, vocabulary, and working with young ELLs. I am grateful to Dr. Tong for helping solidify my understanding of statistics in terms of this dissertation and beyond. I appreciate Kristin and Howard for all that they taught me, and the anxieties they assuaged.

I cannot imagine what I would have done without Ana. She was my emotional support throughout this process. Ana was a constant reminder of the importance of inviting God into this process. It seems that when she was up, I

was down, and when I was up, she was down; we were a perfect balance for each other.

Also, much gratitude is extended to Patrick, Lakshmi, Sandra, Joy, Janie, Polly, Rosie, Tiberio, Julie, and the rest of the bilingual office family, for providing encouragement, accountability, or whatever I needed at the time.

TABLE OF CONTENTS

Page

ABSTRACT iii

DEDICATION v

ACKNOWLEDGEMENTS vi

TABLE OF CONTENTS viii

LIST OF FIGURES xiv

LIST OF TABLES xv

CHAPTER

I INTRODUCTION 1

Lagging in Educational and Financial Attainment 2
Lack of English Proficiency Hinders Academic Success 2
Statement of the Problem 3
Statement of the Purpose 9
Research Questions 11
Significance of the Study 12
Definitions of Terms 15
English Language Learners 16
Assessment 16
Oral Proficiency 17
Typical Transitional Bilingual Education (TBE-C) Model 17
Typical Structured English Immersion (SEI-C) Model 18
Enhanced Transitional Bilingual Education (TBE-E) Model 18
Enhanced Structured English Immersion (SEI-E) Model 19
L1 19
L2 20
Random Effects 20
Fixed Effects 20

CHAPTER Page

Mixed-effects Model 20
Theoretical Framework 20
ELLA Curriculum and STELLA Intervention 20
Activity Structures 22
Language Content 22
Language of Instruction 23
Language Mode 24
Language 24
Communicative Language Ability Model 25
Second Language Acquisition 30
Limitations 31
Delimitations 32
Assumptions 35
Organization of the Study 36

II REVIEW OF THE LITERATURE 37

Introduction 37
Story Retell 38
Story Retell Defined 39
Story Retelling and Reading Comprehension 40
Story Retell and Oral Language 43
ELLs and Story Retell 43
Vocabulary 53
Vocabulary and Reading Comprehension 54
Vocabulary Indirect and Direct Instruction 54
Word Knowledge 56
Assessment Perspectives 57
Assessment Dimensions 57
Choosing Words to Teach 60
Vocabulary Word Levels 60
Vocabulary and ELLs 61
Oral Proficiency 67
Oral Proficiency and Reading 69
Discrete Elements of Oral Proficiency 70
Vocabulary 70
Grammar 72
Syntax 74
Semantics 76
Oral Proficiency Measures for ELLs 76
Curriculum-based Assessment 78

CHAPTER Page

Purpose of Tests 79
Purpose of CBA 79
Standardized and Commercial Tests 80
Quality CBA 81
Teachers and CBA 82
Assessing ESL and Young Children 83
Considerations in Language Testing 84
Psychometric Considerations 84
Validity 84
Reliability 85
Rating Scales 86
Criterion-referenced and Norm-referenced Rating Scales 86
Scoring 87
Rating Scale Problems 88
Raters 88
Teachers as Raters 89
Conclusion 90

III METHODOLOGY 91

Research Design 92
Setting and Participants 92
Setting 93
Participants 95
Procedures 97
Program Intervention 98
Control Structured English Immersion 98
Control Transitional Bilingual Education 99
Experimental Structured English Immersion 99
Experimental Transitional Bilingual Education 100
STELLA Intervention 100
STELLA Oral Proficiency 101
STELLA Vocabulary 102
STELLA Comprehension 103
Instrumentation 104
Development of the S4 104
Project STELLA Vocabulary Fluency Measure 106
Characteristics of Effective Language Instruments 106

CHAPTER Page

S4 Scale Descriptors 107
Iterations of the S4 109
Initial S4 Development 109
First Iteration of the S4 110
Second Iteration of the S4 111
Third Iteration of the S4 112
Fourth Iteration of the S4 113
Woodcock Language Proficiency Battery - Revised 113
Relevant Subtests 114
ITBS 115
Relevant Subtests 115
Research Questions 116
Data Collection 117
Assessment Schedule 117
Commercial Measures 118
Curriculum-based Assessment 118
Data Analysis 121
Summary 122

IV RESULTS 123

Data Exploration 125
Results for Research Question 1 131
Validity 132
Reliability 137
Intrarater Reliability 138
Interrater Reliability 141
Summary 147
Research for Research Question 2 147
Summary 151
Results for Research Question 3 151
WLPB-R Subtests as Covariates of S4 153
Testing for Random and Fixed Effects of Campus and Teacher 155
Testing the Full Model with S4 158
Testing the Full Model with WLPB-R Subtests 159
Effect Size and Summary 165

CHAPTER Page

V SUMMARY, DISCUSSION, IMPLICATIONS RECOMMENDATIONS AND CONCLUSIONS 169

Summary of the Study 169
Discussion by Research Question 171
Research Question 1 171
Validity 172
Reliability 175
Research Question 2 177
Research Question 3 179
Limitations 184
Implications for Practice 187
Utility 187
Cost 189
Recommendations for S4 189
Validity 190
Reliability 191
Intrarater Reliability 192
Interrater Reliability 193
Concluding Remarks 193

REFERENCES 196

APPENDIX A 219

APPENDIX B 220

APPENDIX C 221

APPENDIX D 222

APPENDIX E 223

APPENDIX F 224

APPENDIX G 225

APPENDIX H 227

APPENDIX I 229

APPENDIX J 231

Page

APPENDIX K 233

APPENDIX L 235

APPENDIX M 251

VITA 261

LIST OF FIGURES

FIGURE Page

1 Components of Communicative Language Ability 26

2 Components of Language Competence 28

3 Frequency Graph of the S4 with Normality Curves 129

LIST OF TABLES

TABLE		Page
1	Ethnic Distribution of Students in District T and in Texas	94
2	Kindergarten ELL Participants in Project ELLA, 2004-2005	97
3	Leung and Pikulski's Descriptors for Vocabulary Analysis	107
4	Eller, Pappas, and Brown's Descriptors for Vocabulary Analysis	108
5	Assessment Schedule for Project ELLA, 2004-2005	117
6	Descriptive Statistics for the Pretest Scores on WLPB-R Subtests	126
7	Descriptive Statistics for Posttests and Subtests Measures	127
8	Means of WLPB-R Pretest and Posttests by Group	130
9	Cohen's d and Effect Size for the WLPB-R Pretests and Posttests by Group	131
10	Table for Interpreting Correlation Coefficients	133
11	Correlation Coefficients for the S4, WLPB-R and ITBS ITBS Posttest Measures	135
12	Correlation Coefficients for the S4 with WLPB-R and ITBS by Group	136
13	Intrarater Reliability Correlation Coefficients, Effect Sizes, and Percent Agreement	139
14	Table for Interpreting Kappa	140
15	Interrater Reliability Statistics for First Iteration	143
16	Interrater Reliability Statistics for Second Iteration	145

TABLE Page

17 Interrater Reliability Statistics for Third Iteration 146

18 Interrater Reliability Statistics for Teacher Raters for Fourth Iteration 149

19 Time in Minutes Expended by Teachers Using the S4 Manual and Self-training Materials 150

20 Descriptive Statistics for Pretest on WLPB-R (Subtest Measures) 154

21 Information Criteria Used to Determine Model Fit with DV S4 157

22 Estimates of Covariance Parameters for Campus and Teacher (Classroom) for S4 158

23 Type III Fixed Effects with DV S4 159

24 Type III Fixed Effects with DV WLPB-R (Post Picture Vocabulary) 161

25 Type III Fixed Effects with DV WLPB-R (Post Listening Comprehension) 162

26 Estimates of Covariance Parameters for Campus and Teacher (Classroom) for WLPB-R Post Listening Comprehension 163

27 Type III Fixed Effects with DV WLPB-R (Posttests Vocabulary Analogies) 164

28 Estimates of Covariance Parameters for Campus and Teacher (Classroom) for WLPB-R Post Vocabulary Analogies 165

29 Summary of Effects by Measure 167

30 Means on the S4 in Each Group 168

TABLE Page

31 Traditional and Alternative Assessment 174

32 Summary of the Type III Effects of the S4 181

CHAPTER I

INTRODUCTION

According to the National Center for Education Statistics (2005), there has been an increasing number of young children in public schools who are both linguistically and culturally diverse. In 2004 (the year in which data were collected), the national percentage of children who were 5-years-old and under was 56%; White nonHispanic, 21.8%; Hispanic, 14.5%; Black nonHispanic, 4.2%; Asian or Pacific Islander, 0.9%; American Indian or Alaskan Native and other, 2.7%. Nationally, the largest minority student population is Hispanic. Of the students classified as English language learners (ELLs) or Limited English Proficient (LEP), the majority (79%) speak Spanish as their first language (L1) (Kindler, 2002).

The data for my study were collected in Texas where Hispanics have not only represented the largest minority, but they have comprised the largest segment of the student population as a whole: 45.3%. Whites represent 36.5%, African-Americans represent 14.7%, Native Americans represent 0.3%, and Asian or Pacific Islanders represent 3.1% of the student population (Texas Education Agency, 2006a). In Texas, 93.4% of the ELL population was comprised of Spanish speakers. Steve Murdock (2006), former Texas state demographer, projected that by the year 2040 the population in schools will be 66.3% Hispanic, 19.9% Anglo, 8.3% Black, and 5.5% other.

This dissertation follows the style of *Journal of Educational Psychology*.

Lagging in Educational and Financial Attainment

The NCES (2007) revealed that language minorities trailed behind their English-speaking counterparts in high school completion, enrollment in postsecondary institutions, and overall educational attainment. The disparity was prevalent with Spanish-speaking minorities who had low English proficiency. On the other hand, language minorities who spoke English well, manifested no detectable difference in college enrollment and educational attainment (U.S. Department of Education, 2004). College enrollment and educational attainment are closely tied to earning potential. According to Murdock (2006), in Texas in 1999, the average household income of Hispanics was the lowest among racial groups in the state of Texas: $29,873, in comparison Anglos and Asians earned $47,162 and $50,049, respectively. This disparity can be attributed to the high school dropout rate of Hispanics, which was at 50%. None of the other groups had above 25% of their population drop out of high school. By 2006, only 10% of Hispanics were attending institutions of higher education (Murdock, 2006). These numbers are disconcerting because the largest segment of Texas's population is projected to have a low-level of educational and financial attainment (Murdock).

Lack of English Proficiency Hinders Academic Success

Lack of English proficiency has hindered the academic success of many ELLs. Researchers García-Vásquez, Vásquez, López, and Ward (1997) found significant connections between English proficiency, standardized achievement

scores, and grade point averages. However, research over the last 20 years has not focused on English oral language outcomes (Genesee, Lindholm-Leary, Saunders, & Christian, 2004). There continues to be a paucity of empirical research that informs instruction, in terms of oral language development for academic purposes (Saunders & O'Brien, 2006). Markedly, the paucity of research presents a concern because oral language competence was found to relate to reading outcomes and achievement among native English students (Biemiller, 2003). There is an emerging contribution of research in the field by some researchers (Tong, 2006; Tong, Irby, Lara-Alecio, & Mathes, 2008a; Tong, Lara-Alecio, Irby, Mathes, & Kwok, 2008b) that has as its foci the improvement of oral language development, language acceleration, and effects of program types on language instruction and acquisition. However, more research is needed to continuing adding to the knowledge base of oracy and vocabulary development for ELLs in order to enhance the academic achievement of ELLs.

Statement of the Problem

What does integrating and examining the corpus of research on vocabulary acquisition, oral language proficiency, and curriculum-based assessment reveal in terms of enhancing the academic performance of ELLs? Among the many foci of educating ELLs, language proficiency should be primary because learning was deemed, for over 40 years, to be predominately a language-based activity (Britton, 1970), and brain research substantiated that language development is the foundation for educational achievement in the

elementary grades (Watson, Layton, & Abraham, 1994). A convergence of evidence suggested that ELLs require instruction to develop higher order thinking skills and boost oral language ability via vocabulary knowledge of Tier II words (Beck, McKeown & Kucan, 2002) (i.e. words which provide a framework for elaborate speech). It is erroneous to assume that ELLs would *catch-up* academically and linguistically to English-proficient peers, because as Cummins (1996) articulated that every year native-English speaking students gain sophisticated vocabulary, grammatical knowledge, and increase their literacy skills; they become moving targets. Native-English proficient students do not halt their academic progress in order to allow ELLs to attain the same level of academic and linguistic competence. Also, ELLs have fallen behind native-speakers in content area instruction because they have devoted substantial time to learning English, while native-speakers advance in content instruction (Thomas, 1992). Carlo, August, McLaughlin, Snow, Dressler, Lippman, Lively, and White (2004) reinforced that there is a need to close the gap between native-English students and Spanish-speaking ELLs students. Unfortunately, trying to bring ELLs on par with native-English speakers has been a challenge for many teachers evident in that teachers have expressed confusion about how to best support the English oral language development of ELL students (Gersten & Baker, 2000).

In addition, assessing the oral language proficiency of students has not been without problems. Saunders and O'Brien (2006) found that most oral

proficiency instruments have not been subjected to rigorous evaluation. Specifically, they found that classification cutoffs for proficiency levels varied from one test to the next, normative results were problematic depending on the match between examinees and the norming group, and nonproficient classifications were inaccurate (Saunders & O'Brien, 2006). Namely, tests (i.e. general, commercial, standardized tests) were found to be independent from the curriculum being used. These extant commercialized tests have not provided direct help in meeting the daily curricular or instructional demands placed on teachers and students (Hargis, 1995).

According to Muter and Diethelm (2001), some research studies have shown that vocabulary either influenced or correlated positively with ELLs' early reading-related skills including phonological, orthographic, and morphographic processes. Some researchers claimed that vocabulary knowledge was a causal determinant of differences among students' reading ability and comprehension (Stahl & Fairbanks, 1986; Stanovich, 1986). Garcia (1991) found that unfamiliar English vocabulary was a major linguistic factor negatively affecting the reading test performance of Latino/a students. ELLs who have had slow vocabulary growth were at a disadvantage in terms of textual comprehension, as compared to their monolingual native-English speaking peers, and they have been at greater risk of being labeled as learning disabled, when in reality, the source of the problem was poor comprehension because of vocabulary knowledge limitations (August, Carlo, Dressler, & Snow, 2005). Although, vocabulary

development has been established as crucial for academic success, printed words and spelling continue to supersede vocabulary in reading instruction (Biemiller & Slonim, 2001). Vocabulary has received less attention because of the uncertainty of how to assess vocabulary (Biemiller, 2004).

In considering oral proficiency and vocabulary development of ELLs, another dimension to the problem surfaced for students and teachers. The problem was reflected in the exigencies placed by accountability legislation such as the *No Child Left Behind Act* (U.S. Department of Education, 2003). Legislation required that schools provide scientifically-based instruction. To identify scientifically-sound and efficacious programs for ELLs, it became important to explore student performance and achievement across program models, using data from interventions and assessments. Two of the most common models used with ELLs which have been continuously surrounded by polemics have been Transitional Bilingual Education programs and Structured English Immersion programs (Crawford, 2000). In 1991, a congressionally-mandated study, Longitudinal Study of Structured English Immersion Strategy, Early-exit and Late-exit Transitional Bilingual Programs for Language Minority Children (also known as the Ramírez report), examined the effectiveness of two alternative program models: Transitional Bilingual Education (early-exits) and Transitional Bilingual Education (late-exit) with Structured English Immersion. In the study, the students in TBE outperformed and were at an advantage over students in SEI programs (Ramírez & R.T. International, 1992). However, some

researchers did not support the findings of this study because they claimed the study was inchoate for sundry reasons. For example, Baker (1992) relayed that some weaknesses of the Ramírez Report included (a) a weak theoretical framework; (b) the Hierarchical Liner Model Analysis (HLM) analysis in the study showed that there were strong effects for bilingual education within the first year of schooling but that effect decreased in subsequent years; (c) the academic performance effects found in the study were tied closer to school and district effects, as opposed to program (immersion, early-exit, and late-exit) effects; (d) the Trajectory Analysis of Matched Percentiles (TAMP) analysis could not isolate where growth differences were occurring (district, students, school, or program); and (e) normative comparison were made and these were deemed inadequate because they represent student growth and not program effects. Another researcher, Rossell (1992) stated that the concerns with the Ramirez included: (a) the lack of comparison of early-exit, late-exit, and immersion programs to each other; (b) data for fifth- and sixth- grade immersion and early-exit participants were not collected and included in the analysis; (c) programs (immersion, early-exit, and late-exit) were used nominally and the programs were not thoroughly defined and anchored through the percentage of English used in each program; and (d) the researchers did not study pull-out English as a Second Language (ESL) Models. On the other hand, Collier (1992) and Thomas (1992) concurred with the findings of the Ramírez Report despite

questions about the methodology of the congressionally-mandated longitudinal study or the political issues that might have influenced the conclusions.

In some research studies, as was the case in the Ramírez Report, the tests administered to the participants, which provided data for analyses in the study, were standardized and commercial tests. However, it is important to consider that standardized, norm-referenced, and commercial tests should not be the sole means for assessing student performance and academic achievement, because commercial tests provide an overview of the school and they do not show progress in terms of what the teachers are teaching (Elford, 2002). Assessment and observation in the classroom should have instructional usefulness by focusing on language learning that is relevant to instruction (Genesee & Upshur, 1996).

Furthermore, curriculum-based measures facilitate repeated administrations of the test in order to better monitor student progress on a given skill or skill set overtime and repeated administrations of equivalent measures, are difficult to do with standardized testing (Roberts, Good, & Corcoran, 2005). Standardized measures can be expensive, thus limiting their utility (Genesee & Upshur, 1996). According to Genesee and Upshur (1996), one assessment tool that is time-consuming to create but allows for repeated use is a rating scale. Rating scales can be instrumental in sharing information with parents, other teachers or specialists, as well as for formal grading because they provide information on a relatively observable and specific aspects of language use

(Genesee & Upshur, 1996). Based on the corpora of research on vocabulary, oral proficiency, and curriculum-based assessment it became evident that in order to contribute to the academic success of ELLs it was important to create a curriculum-based instrument to measure vocabulary and oral proficiency in a rating scale format that teachers could easily and repeatedly use. My study is an attempt at creating such an instrument.

Statement of the Purpose

The first purpose of my study was to develop and validate a curriculum-based assessment measure for expressive vocabulary and oral proficiency: Semantic and Syntactic Scoring System (S4) (see Appendix A). The S4 was used to analyze the responses provided on the Project STELLA Vocabulary Fluency Measure (see Appendix B) by the kindergarten students in the large-scale project English Language and Literacy Acquisition (ELLA) (Lara-Alecio, Irby, & Mathes, 2003) . As part of the development and validation process teacher utility was also considered. The second purpose of my study was to use the S4 instrument and other commercial measures such as the language and vocabulary subsections of the Woodcock Language Proficiency battery-Revised (WLPB-R) and language and vocabulary subsections of the Iowa Test of Basic Skills (ITBS) to compare the performance of students who partook in instruction under the two most common bilingual education models: Transitional Bilingual Education (TBE) and Structured English Immersion (SEI), with control and experimental treatments for each under the Project ELLA.

The data for my study were from a longitudinal, five-year, federal experimental research project: English and Literacy Acquisition (ELLA) (R305P030032) (Lara-Alecio et al., 2003). The overall goal of ELLA was to examine the impacts of two different programs on the performance of Spanish-speaking English language learners in grades K to 3 by developing, implementing, and evaluating two research-based models of instruction: a structured English immersion program and a transitional bilingual education program. The intent of the investigators in Project ELLA was to determine interventions that would improve English proficiency and reading achievement. The first year of the intervention had a sample of 1152 kindergarten Spanish-speaking ELLs. These students received services in two program models with a control and experimental group in each. The groups were as follows: (a) control Transitional Bilingual Education (TBE); (b) experimental TBE; (c) control Structure English Immersion (SEI); and (d) experimental SEI.

To ameliorate the deficits in oral language, vocabulary, comprehension, and lack of higher-order thinking opportunities for second language learners, among other interventions, the *Story-retelling and higher order Thinking for English Literacy and Language Acquisition (STELLA*) (Irby, Quiros, Lara-Alecio, Rodríguez, & Mathes, 2008; B. J. Irby, Lara-Alecio, R., Quiros, A. M., Mathes, P. G., & Rodriguez, L., 2004) was created for the TBE and SEI experimental groups in Project ELLA. The students in the experimental programs were assessed using the Project STELLA Vocabulary Fluency Measure which was

modified from the DIBELS measure Word Use Fluency – Grades K and First (Good & Kaminski, 2002; Good et al., 2003). The Project STELLA Vocabulary Fluency Measure required student to produce a sentence for 18 curriculum-based vocabulary words. The crux of this study was contributing to the assessment of the data provided by the students on the Project STELLA Vocabulary Fluency Measure by creating and validating a companion rubric for that assessment.

Research Questions

The following three questions guided my study:

1. To what extent can a curriculum-based measure be developed and validated to measure oral proficiency and vocabulary knowledge among ELLs who are participating in a controlled oral language development intervention?

2. To what extent can teachers use the Semantic and Syntactic Scoring System (S4) for the STELLA vocabulary fluency measure with minimal training to accurately assess students' vocabulary knowledge and oral proficiency?

3. To what extent does the developed curriculum-based assessment instrument, Semantic and Syntactic Scoring System (S4) differentiate the level of knowledge regarding expressive vocabulary and oral proficiency of kindergarten students participating in the STELLA intervention under two different programs: enhanced Traditional Bilingual Education and the enhanced Structured English Immersion Program in comparison to the Revised Woodcock Language Proficiency battery (WLPB-R) (language and vocabulary subtests)?

Significance of the Study

The significance of my study stemmed from the premise that developing language proficiency enhances academic success for young children (Biemiller, 2003; García-Vázquez et al., 1997). Learning is a language-based activity and provides the foundation for academic success, particularly for elementary children (Britton, 1970; Watson et al., 1994). Second language learners need instruction in oral proficiency, vocabulary acquisition, and higher order thinking skills (Beck, McKeown, & Kucan, 2002). This instruction should be focused, purposeful, and intensive (Carlo et al., 2004), because one cannot assume that ELLs will eventually catch-up to English-proficiency students, for students with English as their L1 are continuously growing academically; they do not cease their academic growth to allow English L2 students to be on par with them (Cummins, 1996). A synthesis of research conducted by Saunders and O'Brien (Genesee, Lindholm-Leary, Saunders, & Christian, 2006) elucidated that there were a myriad of problems surrounding the corpus of oral proficiency research and instruments developed to measure oral proficiency, such as: (a) most research is focused on a single academic school year as opposed to being longitudinal; (b) the research does not inform about the developmental language changes that ELLs undergo from novice to advanced; (c) current literature does not examine the interdependence and simultaneous development of oral language, literacy, and academic skills; and (d) participants in studies need to be more diverse because most ELLs research takes place with young children

of Hispanic backgrounds. Also, the literature seems to focused on language of instruction In addition, current oral proficiency literature does not include studies that address the development of academic oral English and how to accelerate the oral language development of ELLs, nor does it examine the impact of program types on oral language development (Tong, Lara-Alecio, Irby, Mathes, & Kwok, 2008b)

The body of research on vocabulary acquisition and measurement has been and will continue to thrive (Gardner, 2007; Read, 2000). However, research studies in vocabulary are inconsistent with each other for the following reasons: (a) many of the studies do not have well defined vocabulary constructs which can be used for comparison among studies, (b) authors use different perspectives in approaching their research in vocabulary, and (c) researchers treat vocabulary as a separate construct and do not always consider or study the impact of other language factors on vocabulary acquisition (Hiebert & Kamil, 2005; Read & Chapelle, 2001).

My study contributes to the corpus of literature on oral proficiency and expressive vocabulary as an integrated concept, because researchers (Alderson & Banerjee, 2002; Bachman, 1990; Canale & Swain, 1980; Singleton, 1999; Verhoeven & Vermeer, 1992) have concluded that there is a strong case to treat constructs of language in light of language factors. For example, a lexical construct would include vocabulary as an integral component of language (oral proficiency, syntax, semantics, grammar, socio-linguistic factors) and must

be measured as such and not in isolation to other language skills (oral proficiency, syntax, grammar) (Alderson, Banerjee, Bachman, Canale, Swain, Singleton, Verhoeven, & Vermeer). However, researchers, Harley, Allen, Cummins, and Swain (1990, p. 24) stated that “an inherent difficulty in validating models of L2 proficiency is that measures faithfully reflecting a particular construct may not have adequate psychometric properties, while other psychometrically acceptable measures may fall short of representing the construct.” Paulston (1990) succinctly stated that what is needed is qualitative and quantitative approaches in order to understand second language acquisition because quantification and psychometrics are not sufficient in terms of measuring language.

In a National Literacy Panel study, the researchers found that adequate assessments were essential for program placement, tailoring instruction, and evaluating progress of second language learners; however, in the same study the researchers reported that extant assessments were inadequate in providing needed information of language proficiency (August & Shanahan, 2006). Furthermore, the National Literacy Panel (August & Shanahan, 2006) study found that teacher judgment and assessment are significant in the education of language-minority students; therefore, the researchers recommend that additional research explore teacher judgment and assessment further.

Therefore, I worked with two researchers involved in STELLA, Irby and Pollard-Durodola, in developing a curriculum-based instrument, Semantic and

Syntactic Scoring System (S4) (Walichowski, Irby, & Pollard-Durodola, 2007) that was used to measure the construct of oracy and expressive vocabulary, applicable to the vocabulary instruction that children were receiving in the STELLA intervention. The S4 is an instrument that facilitates longitudinal data collection in the primary grades, measures an integrated construct of language through expressive vocabulary, oral proficiency, semantics, and can be used with Spanish-speaking ELLs or ELLs in general. The S4 collects qualitative data and quantifies it. The qualitative data are the sentences that students construct and orally provide using target vocabulary words. Then the S4 provides a 5-point scale which is used to rate word knowledge, semantics, and syntax of the sentences produced; thus, providing quantitative data that can be analyzed and compared to evaluate student progress and performance. The instrument can be used to inform and evaluate vocabulary instruction and to objectify teacher judgment of language proficiency for students.

Definitions of Terms

It is important to establish working definitions for assessment, English Language Learners, the construct of language, when conducting research in the area of vocabulary and oral proficiency. Establishing working definitions facilitates generalizations when generalizations are appropriate. The way one defines assessment, population, and the construct of language has a direct impact on one's research approach and research findings.

Other terms were included in an effort to anchor this study. The following definitions were adopted because they related to my study.

English Language Learners

English language learner (ELL) denotes students who initially learned another language in their home and community before learning English. These students could have been immigrants or U.S. born. The students might have had some knowledge of English, but when they entered school, they were not proficient English speakers. Other interchangeable terms for ELLs are limited English proficient, non-native English speaker, language minority student, English as a Second Language (ESL) student, or bilingual student (Genesee et al., 2004).

Assessment

Bachman (2004) defined assessment as, "assessment can be thought of broadly as the process of collecting information about a given object of interest according to procedures that are systematic and substantively grounded. A product, or outcome of this process, such as a test score or a verbal description, is also referred to as an assessment" (Bachman, 2004). This definition of assessment is important because it atones for the myriad instruments and probes that do or attempt to measure vocabulary knowledge and oral proficiency. Based on this premise of what assessment is, a teacher's observations and perceptions of language ability can be presented as a valid measure of vocabulary knowledge and oral proficiency. This allows for

alternatives to traditional criterion, standardized, and norm-referenced tests. It is important to consider the broad spectrum of assessment possibilities. Perhaps using a combination of assessment alternatives to validate or complement an instrument would be beneficial. Furthermore, according to Boehm (1992), the word *assessment* as it relates to early childhood education is interchangeable with the word *measurement*, which is a procedure that one uses to determine the degree to which a child possesses an attribute of scholarly interest.

Oral Proficiency

Oral proficiency can be considered the set of words that are known because they are spoken and read aloud. Sometimes when referring to oral proficiency, different facets of language are included, as maintained in Hargett's (1998, p. 8) statement: "To be proficient in a second language means to effectively communicate or understand thoughts or ideas through the language's grammatical system and its sounds or written symbols."

Typical Transitional Bilingual Education (TBE-C) Model

As defined in the ELLA project, this program model began in kindergarten with 80% English instruction and 20% Spanish instruction. It gradually increased the amount of English and reduced the amount of Spanish until, by grade three, both languages are spoken 50% of the time. Under this model, the initial goal in kindergarten is to focus on oral language, moving to content instruction in Science and Social Studies in English by grade three. This also was referred to as the control TBE.

Typical Structured English Immersion (SEI-C) Model

As defined in the ELLA project, this program model was taught in district A. In this type of program, all subjects were taught in English. L1 clarifications in Spanish were rarely made.

Enhanced Transitional Bilingual Educational (TBE-E) Model

As defined in the ELLA project, this program model began in kindergarten with 70% of instruction in Spanish and 30% of instruction in English. By third grade, Spanish was decreased to 40% and English was increased to 60%. Kindergarten focused on oral language development and then moved to content instruction in science and social studies by third grade. Teachers used content instruction to improve oracy, literacy, vocabulary, and comprehension. Under this model, teachers participated in weekly staff development opportunities in various areas: (a) enhancing instruction via planning, (b) supporting student involvement, (c) vocabulary building and fluency, (d) oral language development, (e) literacy development, including the use of technology, (f) reading comprehension, and (g) parental support and involvement. The paraprofessionals that work under this model were trained to work with students in an intensive English program.

The major differences between the typical (control) TBE and an enhanced (experimental) TBE were: (a) additional time spent in English language acquisition strategies, (b) ongoing professional development and portfolio assessment, (c) parent training, (d) use of the Traditional Bilingual

Observation Protocol (TBOP) instrument (Lara-Alecio & Parker, 1994) to observe teacher practices and provide feedback, and (e) training paraprofessionals to work with this model.

Enhanced Structured English Immersion (SEI-E) Model

As defined in the ELLA project, SEI-E offered all instruction in English with minor use of L1 clarifications. Under this model, teachers participated in weekly staff development opportunities in the following areas: (a) enhanced instruction via planning, (b) support for student involvement, (c) vocabulary building and fluency, (d) oral language development, (e) literacy development, including use of technology, (f) reading comprehension, and (g) parental support and involvement. The paraprofessionals who work under this model were trained to work with students enrolled in an Intensive English program.

The major differences between the enhanced SEI and traditional SEI were (a) there was additional time devoted to English language acquisition, (b) there was ongoing professional development and portfolio assessment, (c) there was a parent training component, (d) the Traditional Bilingual Observation Protocol (TBOP) instrument (Lara-Alecio & Parker, 1994) was used to observe teacher practices and provide feedback, and (e) paraprofessionals were trained to worked with this model.

L1

L1 refers to the first language acquired or ‘mother-tongue’, which in this study is Spanish.

L2

L2 refers to the second language or target language, which in this study is English.

Random Effects

"A random effect presumes a representative sample of levels from the more numerous potential levels on the way, along with interest in generalizing from the sample levels to the population of all possible levels" (Thompson, 2006, p. 346).

Fixed Effects

"A fixed effect occurs when we use all conceivable levels of a way, or (b) we do not want to generalize beyond the levels we actually employ" (Thompson, 2006, pp. 345-346).

Mixed-effects Model

"A mixed-effects model occurs when at least one omnibus hypothesis is treated as a fixed effect, and at least one omnibus hypothesis is treated as a random effect" (Thompson, 2006, p. 346).

Theoretical Framework

ELLA Curriculum and STELLA Intervention

The data for this study were from Project ELLA (Lara-Alecio et al., 2003) and the STELLA intervention (Irby et al., 2008). It is important to examine the theoretical premise for ELLA and STELLA in order to elucidate the foundational theory from which the data were derived for this study. The theoretical

foundation for Project ELLA and STELLA was the Four Dimensional Transitional Pedagogical Theory (Lara-Alecio & Parker, 1994). The four dimensions of the model are academic structures, language of instruction, language content, and communication mode. Lara and Parker developed this theory in response to an evaluation of extant theory and principles in the field of bilingual education. They surmised that most of the theories in the field were not emerging from classroom settings and were not translated into principles that could directly be applied and impact praxis in classrooms.

The general principles they noted from the corpus of bilingual theories were as follows: “Provide an emotionally supportive environment; emphasize quality of social interaction between teacher and student; ensure ‘bilingual’ status is not considered a disability; provide quality social interactions between teacher and student; provide multiple-modality interactions with student; incorporate minority students’ culture in teaching; guide and facilitate rather than control student learning; encourage student talk and independent learning; structure activities which facilitate quality interactions; encourage community participation in schooling; promote student intrinsic motivation; teach ‘meaningful’ content; develop prior competency in the home language; and continue to develop competencies in both languages” (Lara-Alecio & Parker, 1994, pp. 119-120). Furthermore, Lara and Parker (1994) developed a classroom observation instrument, Transitional Bilingual Observation Protocol (TBOP) to evaluate classroom instruction based on the Four Dimensional

Transitional Pedagogical Theory. The TBOP has been validated and applied in second language acquisition classroom settings (Breunig, 1998; Meyer, 2000).

Activity Structures. In the model, activity structures is defined as "…relatively stable, recurring periods of activity, each with a recognized purpose and opportunities for communication. Communication which is expected, appropriate, and fostered in one activity structure may be inappropriate and discouraged in a second" (Lara-Alecio & Parker, 1994, p. 121). Operationally, activity structures are defined as "(a) type of teacher behavior (e.g. directing, leading, evaluating, and observing), and (b) the expectation for student responding (e.g. listening, performing, discussing, asking questions, answering questions, cooperative learning). A few classroom activity structures (e.g. time spent disciplining, transitioning between classes) are considered non-academic. The TBOP evaluates activity structures in pairs as one teacher behavior and one student behavior (Lara & Parker).

Language Content. In the Four Dimensions Transitional pedagogical Theory, Lara and Parker revisited Jim Cummins's (1986; Cummins, 1996) concepts of Basic Interpersonal Communication Skills (BICS) and Cognitive Academic Language Proficiency (CALPS). Within these two levels Lara and Parker added some additional levels to narrow the gap in the continuum between BICS and CALPS and to apply this concept of language competence to encompass a greater range of classroom discourse. The additional elements included are:

1. Social Routines (i.e., social exchanges and conversation); 2. Academic Routines (i.e., preparing for recess, returning books, learning strategies, handing in assignments, structuring homework); 3. Light Cognitive Content (i.e., current events, discussion of the school fiesta, multicultural education issues, also repetitive drill or skills practice); and Dense Cognitive Content (i.e., new content-area information, conceptually loaded communication with specialized vocabulary and procedures) (Lara & Parker, 1994, page 122).

Furthermore, the model goes beyond generalized developmental sequences and looks at language development as something incremental which changes from time to time and fluctuates based on the activity structure. The model also addresses the cross-linguistic impact of L1 on L2 and L2 on L1, because it allows the evaluation of proficiency based on differences in language activities (social routines, academic routines, light cognitive content, dense cognitive content). This is more sensitive to language development variation instead if assuming that language in (L1 and L2) grown in a rigid and sequential way: such as needing L1 fluency in a particular area before L2 fluency can be had in the same area.

Language of Instruction. Language of Instruction is the third dimension of the Four Dimensions Transitional Pedagogical Theory. This component facilitates the language of instruction decision (determining whether to use L1 or L2 to teach a subject area). In the model Lara and Parker (1994, p. 124) presented the following as combinations of native language and English:

> I. Content presented in L1 (indicates Spanish-only instruction, a beginning point for students with very low English proficiency); II. L1 Introduces L2 (indicates instruction primarily in L1, but additionally, English vocabulary is taught for key ideas, concepts, and procedures); III. L2 Clarified by L1 (Indicates instructional primarily in English, but with

L1 used as a ‘back- up’ as needed, to ensure understanding); and IV. Content presented in L2 (indicates English-only instruction, the goal).

Language Mode. The final component, addresses the limitation posed by mode (reading, writing, and verbal expression) on language facility. The premise of the model is that English proficiency may vary based on the mode that is used and that each mode should be fostered to grow irrespective of English proficiency in the other modes. Lara and Parker (1994, p. 124) explained this as “This may mean that students are permitted to produce an essay exam in L1 on a difficult topic following a lecture presented in English. It may mean that students are expected to read an assignment in English, but follow-up discussion is conducted in L1.”

Language

Language is a complex communication system, which has been analyzed on several levels: phonology, syntax, morphology, semantics and lexis, pragmatics, and discourse (Mitchell & Myles, 1998). Past and current research on language and language measurement has offered many confounding principles. In respect to my study, two antipodal views were considered: (a) that linguists should study meaning and that form and meaning are inseparable which makes lexis and grammar interdependent (Firth, 1957; Halliday, 1985; Sinclair, 1966), and (b) that there is a split between language competence and performance, thus linguists and language researchers should study and model underlying language competence, rather than the performance data of actual,

produced, utterances (see Chomsky, 1957; Chomsky, 1965). But even Chomsky (1957) had tenuously concurred that "the fact that correspondence between formal and semantic features exists…cannot be ignored" (1957, p.102). My study's premise was grounded in the first view and attempted to evaluate language proficiency in terms of the interdependence of oracy and expressive vocabulary.

Since the 1980s, language testing specialists have begun to include theoretical frameworks on language proficiency to guide the methods and technology used in researching and assessing language proficiency (Bachman & Clark, 1987). For this study, I proposed Bachman's (1990) model of communicative language ability as an umbrella theory; because, it is compatible with the premise that language, in terms of lexis and grammar, is interdependent and should be measured as such. Furthermore, Baker (endnote 1996, p. 30) proposed the Bachman's model of language competence as an possible "integrating consideration of the themes of the definition and measurement of bilingualism."

Communicative Language Ability Model

Bachman's communicative language ability model (1990) evolved from the work of several linguists such as Hymes (1972), Munby (1978), Canale and Swain (1980), Savignon (1983), and from Bachman and Palmer's (1982) earlier, empirically-based work in which they explored the construct validity of tests that purported to measure components of communicative competence. Bachman

and Palmer (1982) conducted a study to determine if frameworks of communicative competence proposed by these researchers (e.g., Canale & Swain, 1980; Hymes, 1972; Munby, 1978; Savignon, 1983) had components of language competence that could be defined and distinguished from one another. They empirically studied the construct validity of three distinct traits: linguistic competence, pragmatic competence, and sociolinguistic competence and found distinctiveness. From the linguistic theories and the construct validity analyses, Bachman created a model that contained three interacting components: language competence, strategic competence, and psychological mechanisms as indicated in Figure 1.

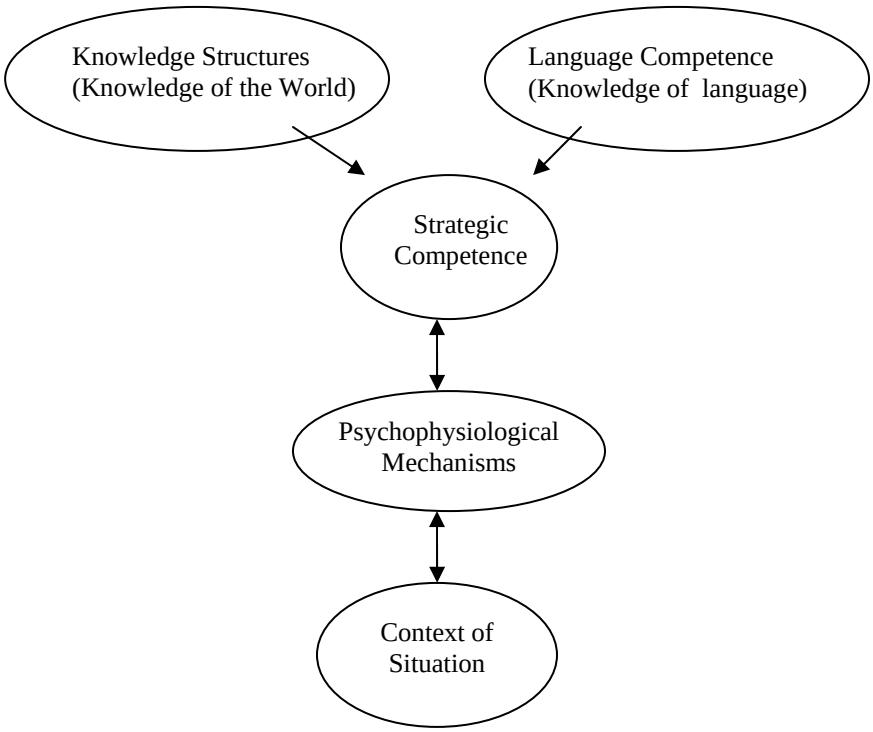

Figure 1. Components of communicative language use (Bachman, 1990, p.85) adapted with permission.

For the purpose of my study the domain of language competence (knowledge of language) was the most relevant. Figure 2, represents a hierarchical view of the components of language competence. Language competence consists of two traits, organizational competence and pragmatic competence. Organizational competence includes grammatical and textual competence. Pragmatic competence subsumes illocutionary and sociolinguistic competence. These components are further broken down to provide a more detailed description of the construct. Bachman developed a diagram, such as the one depicted in Figure 2. Bachman (1990, p. 86), stated that his diagram represented more of a metaphor than an actual model because it captured "the hierarchical relationships among the components of language competence, at the expense of making them appear as if they are separate and independent of each other."

The aspect of the model that was of interest for this study was specifically grounded in the left side of the model, organizational competence; because, "…language assessment should be carried out within the framework which takes the formal properties of language into account" (Pienemann & Johnson, 1987, p.91). Organizational competence deals with the structures formed by formal properties of language. This formal structure is what facilitates the production or recognition of grammatically correct sentences, understanding of their propositional content, and order that helps form texts (Bachman, 1990).

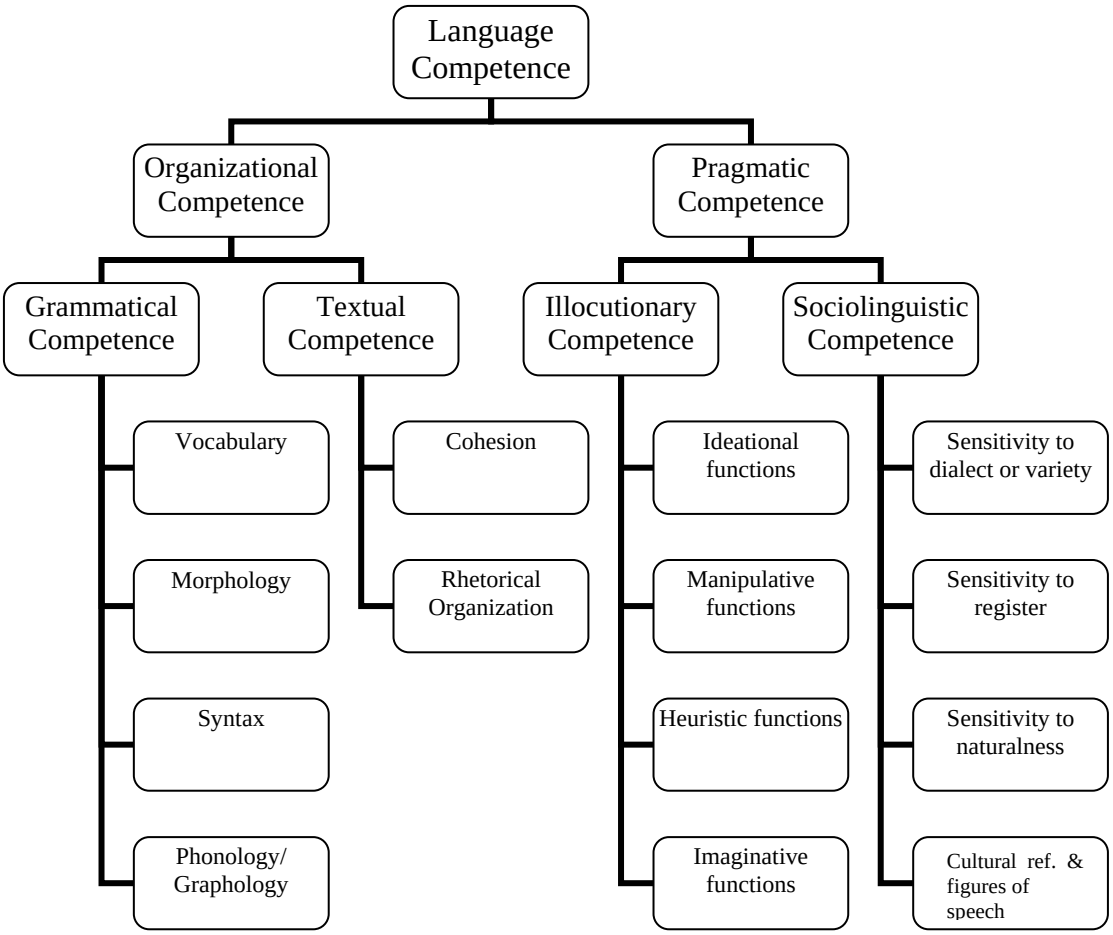

Figure 2. Components of language competence (Bachman, 1990, p.87) adapted with permission.

The Grammatical Competence in Bachman's model includes those competencies involved in language usage: vocabulary, morphology, syntax, and phonology/graphology. Bachman provided the following example (which is similar to what was elicited from students in this study; with the difference being that I provided a target word instead of a picture of grammatical competence:

> Suppose, for example, a test taker is shown a picture of two people, a boy and a taller girl, and is asked to describe it. In doing so, the test taker demonstrates her lexical competence by choosing words with appropriate significations (boy, girl, tall) to refer to the contents of the picture. She demonstrates her knowledge of morphology by affixing the inflectional morpheme (-er) to 'tall'. She demonstrates her knowledge of syntactic rules by putting the words in the proper order, to compose the sentence 'The girl is taller than the boy'. When produced using the phonological rules of English, the resulting utterance is a linguistically accurate representation of the information in the picture. (1990, p.87)

The aspect of Textual Competence includes the knowledge and skills that are needed to join together utterances to form a text (a unit of language) that can be spoken or written. Furthermore, Textual Competence consists of two or more utterances or sentences, which are structured, based on rules of cohesion and rhetorical organization. In this study the participants were kindergarten ELLs and the task that was measured was that of producing usually, a single utterance. However, some of the students provided more than a single sentence; hence, textual competence was also determined to be a relevant domain.

Bachman added to his model in 1996, and it included an affective domain which accounted for the "affective or emotional correlates of topical knowledge" (1996, p.65). Furthermore, this addition to the model dealt with interactions that took place between examiners and examinees, when the examinee was emotionally charged or indifferent about the topic being tested. For this study, the affective factor was not deemed relevant and the original model was most

suited given the construct, age group, and type of assessment that was investigated in this study.

Second Language Acquisition

The Bachman model has and will be found to have flaws, but it is the best that is available (Skehan, 1991). As counter criticism to the flaws of the model, Alderson and Banerjee (2002, p.80) have stated that, “nevertheless, we believe that one significant contribution of the Bachman model is that it not only brings testing closer to applied linguistic theory, but also to task research in second language acquisition (SLA), one of whose aims is to untangle the various critical features of language learning tasks.”

The intention of SLA research is to document and explain the learner’s changing interlanguage, and to do so, researchers need reliable descriptions of language at its various stages of development (Bachman, 1990). The second language acquisition stages of development that were considered for this study were those delineated by Ellis (1985): sequence, order, and rate of development. First, there is the sequence in second language learning (which is the same for children and adults). This sequence is based on a natural, universal, and almost invariable sequence of development. The initial part of this stage is evidenced by the production of simple vocabulary and basic syntax, to the structure and shape of simple sentences, to complex sentences. Order, refers to specific and detailed features of language. These features (such as specific grammatical features, situation specific vocabulary) may be acquired

with minor variations based on the individuals. Finally rate of development, refers to the variations in the speed in which the target language is acquired at a proficient level.

Limitations

Readers should interpret the results of this study with some caution. The sample was homogenous on important variables. The participants were from low-income families in a single community, were ELLs,and they may not adequately represent the population of all kindergarteners from any other given community; therefore limiting generalizability. However, generalizability may be inferred to ELLs with similar characteristics as those in my study.

Second, this study was conducted using only kindergarten data. The kindergarten students were from either TBE or SEI classrooms. Because of traditional program placement procedures the students with higher English proficiency were placed in SEI classrooms and those with lower English proficiency were placed into the TBE classrooms.

Moreover, the Mathew Effect was not accounted for in this study. It is possible that children with initial larger receptive and expressive vocabularies would learn more of the target vocabulary words during the story reading sessions than children with smaller initial vocabularies. Furthermore, students that had a greater oral proficiency had an inherent advantage in this type of measure. A baseline was not taken for each participant to determine beginning vocabulary knowledge levels.

A large large *n,* 905 kindergarten students, was used in this study. There were missing data; therefore for some measures there were as few as 500 or 600 scores. Studies with large *n*'s run the risk of making small differences seem significant (Cohen, 1990). However, this study was an exploratory study and it was important to ascertain how students were performing on the S4 and the other measures used in this study. By using a sub sample of the ELLA sample it would have limited the nuances of student performance across the S4 and other measures. Power analysis using sample size calculation statistical software (Raosoft, n.d.) estimated that a sample of 270 would be needed for a confidence interval of 95% and a sample of 384 would be needed for a confidence level of 99%. Either of these sample sizes are considered large in themselves and would be subject the risks of inflating significance as stated by Cohen. Therefore, the entire sample was used because the disadvantage of inflating significance was not greatly changed by using the entire n in comparison to a subsample; but the greater quantity of data allowed for ascertaining nuances of student performance on the S4, in particular, and also across the other measures.

Delimitations

In Texas, random placement of students into treatments was not permitted, so the data acquired through Project ELLA should be treated as experimental/quasi-experimental. The principal investigators in ELLA employed

a robust matching technique to ensure comparability of students in each of the four program types.

My study used one year of data and did not have longitudinal approach. Also this study did not make any attempt at normative comparisons. Three decades ago bilingual education researchers amassed research evidence that demonstrated the importance of studies using long-term assessment (from 4 - 5 years) to best understand students' second language performance (Thomas, 1992). In the Ramírez report (Ramírez & R.T. International, 1992), they were not able to attribute program differences (identify them) until the fourth year, and more data were needed in subsequent years. The same report stated that a direct comparison between language minority performance and native speaker performance in academic achievement provides better information (Ramírez & R.T. International).

The quality of the recordings for the Project STELLA vocabulary fluency measure were generally good. However, there were three tapes (43 students) that had background sound interference. There were also a few occasions in which students (10 participants) spoke too low to have been adequately recorded. If recordings were not audible for a given student or students then those data were not used. There were children (53) that were absent and never received a make-up session for this probe. In all, 10% of the student probes could not be transcribed for scoring.

The data that were used were all recorded and then they were transcribed at a later date for analysis. No transcriptions took place during student testing. Therefore, it is important to note that transcribing during testing and transcribing recorded data after testing can produce differing results. Nambiar and Goon (1993) evaluated studies that used raters for data collection. They compared, the rating of audio-recordings of speaking performances with ratings of live performances and found that raters underestimate the scores of more proficient candidates when they only have access to audio data.

In my study, poor recordings that could not be deciphered were not used. It is expected that some recordings had better audio quality than others, due to testing location acoustics, background noise, proximity of the student to the recorder, and strength of the battery in recorder . It is important to note that some researchers believe that the quality of a recording can influence a rater's judgment. For example, McNamara and Lumley (1997) found that poorly recorded performances tended to be judged more harshly and the interlocutor was deemed less competent; however, when the recordings were of better quality then the interlocutor was judged with greater leniency. Reed and Cohen (2001) advised that careful selection of raters, training of raters, and clear assessment procedures should ameliorate external influences in rating and I applied those suggestion in my study.

Assumptions

As part of this study, it was assumed that the teachers and paraprofessionals knew the STELLA curriculum well and had fidelity in implementing it. Under the ELLA grant there was a STELLA coordinator, curriculum coordinator, and the principle investigators that provided training and fidelity checks for the curriculum. In general, teachers in the experimental classrooms, engaged in biweekly professional development. The paraprofessionals received training once per month. Furthermore, the teachers involved in STELLA were observed and evaluated once per month while teaching to measure curriculum fidelity using the Teacher Observation Protocol, originally the Transitional Bilingual Observation Protocol (Lara-Alecio & Parker, 1994). Therefore, the assumption that the curriculum was followed is a reasonable assumption.

It is also assumed that all testing was conducted in accordance with the testing procedures and manuals for each test and that there was fidelity with each administration. In the ELLA grant there was an assessment coordinator and principle investigators that provided training for the bilingual paraprofessionals and districts substitutes on each instruments' assessment procedures. Each test administrator was given the opportunity to practice giving the assessments. Once the test administrators demonstrated proficiency then they were allowed to test students for data collection purposes.

Organization of the Study

Chapter I of this study contains the Introduction of the study and includes the following: Statement of the Problem, Definitions of Terms, Statement of the Purpose, Research Questions, Significance of the Study, Definition of Terms, Theoretical Framework, Limitations, Delimitations, and Assumptions.

Chapter II of this study consists of the Literature Review and includes the following: Introduction, Story Retell, Vocabulary, Oral Proficiency, Curriculum-based Assessment, Considerations on Language Testing, and Conclusion.

Chapter III of this study focuses on the Methodology and includes the following: Development of the S4, Research Design, Sampling, Program Intervention, STELLA Intervention, Instrumentation, Research Questions, Data Collection, setting, research design, instrumentation, intervention procedure, data collection, Data Analysis, and Summary

Chapter IV of this study depicts the Results and includes the following: Data Exploration, Results by Research Question, and Effect Size and Summary.

Chapter V includes a discussion of findings and the following sections are included: Summary of the Study, Discussion by Research Question, , Limitations, Implications for Practice, Recommendations for the S4, and Concluding Remarks.

CHAPTER II

REVIEW OF THE LITERATURE

Introduction

The purpose of this literature review was to examine the research literature on (a) story retelling, (b) vocabulary, (c) oral proficiency, (d) curriculum-based assessment, and (e) psychometric implications associated with an instrument that attempts to assess vocabulary knowledge coupled with oral proficiency.

When available, the literature reviewed pertained specifically to ELLs. There is much commonality between the process of first language development and second language development. For example, Baker (1996, p. 30) discussed Bachman's Model of Language Competence, which is the foundation for the theoretical framework of this study, as a language structure theory that "…provides an integrating consideration of the themes of the definition and measurement of bilingualism." Specifically, language development in terms of first language and second language acquisition is a subconscious process that is innate and that all individuals have in common, which includes oral and written systems that include phonology, vocabulary, morphology, syntax, semantics, pragmatics, paralinguistics, and discourse (Ovando, Collier, & Combs, 2003). Research on bilingualism as a first language (BFLA), which is when a child learns two languages from birth, reifies that many of the language concepts that are used in first language development are similar to the process

for second language development and this is often discussed as a unitary language system hypothesis (Genesee & Nicoladis, 2007). Therefore, to gain a panoptic understanding of the concepts relevant to this study, it was important to include literature that was not particularized only to ELLs.

Various databases were used to compile information for this literature review. They were as follows: Academic Search Premier, EBSCOhost, Google Scholar, Educational Resources Information Center (ERIC), PsycINFO, ProQuest, Wilson Web, Center for Research on Education, Diversity, and Excellence (CREDE), the Northwest Regional Educational Laboratory, JSTOR, and World Cat. The Boolean connections and variations of key terms used were as follows: story retelling, language proficiency, oral proficiency, oral proficiency measure, oral language development, vocabulary and oral proficiency, and expressive vocabulary, curriculum-based assessment, curriculum based assessment, alternative assessment, classroom-based assessment, internal assessments. Furthermore, these Boolean terms were used in conjunction with English Language Learner, second language learners, second language acquisition, bilingual, and ESL where appropriate to narrow searches. In addition, the reference lists of the studies reviewed were used to identify other important publications to review.

Story Retell

The kindergarten participants of this study partook in a reading intervention, *S*tory retelling and higher-order *T*hinking for *E*nglish *L*iteracy and

*L*anguage *A*cquisition (STELLA) (Irby et al., 2008). STELLA was a scripted, five-day-cycle, 40-minute, structured and interactive story reading pedagogical literacy intervention. STELLA included sundry educational components such as "… (a) integrated ESL strategies, (b) higher ordered leveled questions, (c) academic vocabulary in the content area of science which was explicitly and implicitly taught, (d) opportunities for students to practice language through retelling, and (e) training for the teachers on a biweekly basis" (Irby et al., 2008, p.2). STELLA was systematically developed as an intervention in the ELLA project (Lara-Alecio et al., 2003) for enhancing oral language, vocabulary, comprehension, and higher-order thinking for the students in the experimental Structured English Immersion (SEI) group and the experimental Transitional Bilingual Education (TBE) group. Therefore, in this review, I have included the literature on story retell as it pertained to STELLA and in general. There was a paucity of story-retell literature that dealt with second language learners; thus, it was important to look at the body of research as a whole.

Story Retell Defined

In story-retell tasks, test takers hear or read a story and then retell it. This type of activity can fulfill several objectives: such as listening comprehension, production of oral discourse features, communicating sequences, communicating relationships of events, stress and emphasis patterns, expression, fluency, and interaction with the hearer (Brown, 2004). Some researchers define story retell in the context of cooperative learning groups

(Slavin & Madden, 2006) or peer-assisted groups (Saenz, Fuchs, & Fuchs, 2005).The designers of the STELLA intervention defined structured story retelling as a strategy that ”involves story reading that is systematically planned and scripted to utilize research-based learning strategies” (Irby et al.,2008, ¶ 10). For the purpose of this study the definition of story retelling that was espoused was that given by the researchers of the STELLA curriculum.

Story Retelling and Reading Comprehension

Some researchers (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Pickert & Chase, 1978; Roberts et al., 2005) found that story retell superseded other comprehension assessment formats because it provided a larger sample of student comprehension behaviors, facilitated a sense of story structure (Whaley, 1981), and was time efficient (Roberts et al., 2005). By contrast, other formats such as cloze and question-response do not provide as much information and are limited because they solicit responses under designated parameters (Roberts et al., 2005). Morrow (1990) cautioned, that comprehension assessment is defined by the questions asked, and proposed that it should be the child’s response that is the focal point of comprehension evaluation.

Not only have story retells been used as an assessment of reading comprehension, they have also been used as a tool to enhance reading comprehension. Unfortunately, few studies exist on the impact of story retelling on comprehension. One such study, in which retelling was used to improve comprehension, was conducted by Zimiles and Kuhn (1976). This study was

conducted with 576 participants, aged six to eight. The participants were put into two groups. In this study, half of the children were asked to retell a story after it was read and the other half was not. Then a comprehension test was administered and the results of that test indicated that the students who participated in story retelling scored higher than the students who did not retell the story. Students were also assigned to experimental conditions involving intervals of time elapsed after hearing the story. After analyzing the responses to comprehension questions posed after varying intervals of time, the researchers found that the length of the interval had a decisive influence on the participants' recollection of specific details. Brown's research (1975) suggested that children's story comprehension increased when students actively reconstructed a story by thinking about the individual story's events and pictures and arranging them in sequential order.

Pellegrino and Galda (1982), in their study, found that reading comprehension through story retell improved when children retold the story through active involvement and peer interaction, in the form of role plays. Morrow (1985) conducted two studies using story retelling with children that were already independent readers. The purpose of the first study was to determine if comprehension improved for kindergarteners after listening to a story and retelling it, without frequent practice or guidance during the process. The stories were also analyzed for story elements and syntactic complexity of oral language. In this study the experimental group retold the story; whereas,

the control group drew a picture about the story. After having encouraging findings, that story retell was more effective, a second study was executed which evaluated comprehension and other skill areas based on practice and guidance with story retell. In the second study, students practiced and received guidance during the story retelling process and their retelling was analyzed for average length and syntactic complexity. Morrow claimed that the frequent practice and guidance in retelling rather than review or rehearsal were factors that improved comprehension for the participants.

Most story-retelling research was confined to early childhood and young participants. Barnhart's study (1990) described story-retelling behaviors for children slightly older than the ubiquitous story-retell studies with preschool and kinder students. The participants in his study were second-grade children with diverse reading levels. Barnhart observed the relationship between patterns of retelling behaviors and levels of comprehension for students, who were grouped among three levels: above level, at level, and below level. In the study, the students who demonstrated a clear sense of narrative register and a mature sense of story outperformed the other students. Barnhart made the following statement "research with beginning readers suggests that a well-developed set of semantic and syntactic expectations play a crucial role in successful reading" (1990, p.257). Thus, story retells benefited children who were readers not just students who were at a prereading stage.

Story Retell and Oral Language

Several studies have shown that story retells facilitated language growth (Blank & Frank, 1971; Morrow, 1985; Stewig & Young, 1978). Blank and Frank (1971) conducted a study with children ages four to six in which they studied semantics and syntax of children in a story retell task and a sentence imitation task. They wanted to see if the context support of a story retell, as opposed to isolated sentences, altered the reproduction of syntactic structures. They found that as the children in the study became more familiar with semantic content then they began to elaborate more in their responses. Also, they realized that as with the control group, which could have resembled a typical classroom where students listened to a story without any responses required of them, little material was retained. Pickert and Chase (Pickert & Chase, 1978) suggested that story retelling was a better measure of oral language abilities than just observing the student's language production.

ELLs and Story Retell

Few studies have been conducted with reading comprehension of English language learners (Proctor, Carlo, August, & Snow, 2005). However, some correlational studies found that storybook reading was an instructional tool that impacted vocabulary development for ELLs (Justice, 2002).

In a recent study, on oral narrative skills of bilingual children in the UK, Riley and Burrell (2007) affirmed that effective instruction in language and literacy depended upon having knowledge of young children's oral narrative

skills, particularly when working with children of diverse backgrounds. The ability to narrate and report were deemed vital skills and were highly correlated with future reading fluency. This study (n=120) was with children that were participating in a ‘StoryTalk’, an intervention that is focused on expressive skills, specifically including narrative skills. The participants were taking part in English as an Additional Language (EAL) instruction. The intervention participants (n=60) received a weekly, specially-designed, language-enriched instruction by trained volunteers. The students were tested at the beginning of their first school year and then again at the end. The data for this study were part of a larger language study. For this study the data were collected from a practitioner assessment and a standardized psychological assessment of language. In the story-retell assessment, the teacher evaluated the sentence structure, vocabulary, and global judgments: such as organization, description/expression, and content of the story retell that each child produced. The students took the Clinical Evaluation of Language Fundamentals (CELF) (Wiig, Secord, & Sernel, 1992) which was highly correlated with the Story Talk instrument. The correlation statistic for the CELF and Story Talk was not provided. The CELF assesses language; it is an instrument that is administered by psychologists for clinical assessment and research purposes. The instrument is not intended for teacher administration or teacher use. Riley and Burrell mentioned that other researchers such as Gilmore (1998) and MacDonald and McNaughton (1999) found that Story-Retelling assessments tasks have high validity and reliability,

but they do not assess or report on the validity and reliability of their Story-Retell assessment. The teachers in the study were asked about the story-retell assessment in an informal semi-structured interview. Specifically, they were asked “about the ease of administration, quality of texts, how applicable they considered the assessment to the UK context, and the usefulness of the assessment information” (Riley & Burrell, 2007, p. 186). The researchers found that some teachers had problems scoring/assessing and that practitioners would need more training in using the instrument.

Some of the qualitative data from the Riley and Burrell study illustrated that, as a pretest, the story retell was short, included simple sentences, reflected misconceptions, contained irrelevant materials, and reflected misuse of grammar. When the Story Talk instrument was used a year later, as a posttest, students were able retell a story logically, connect phrases, and included a sense of accurate sequence. Vocabulary was also more developed and reflected the use of literary language in the post test administration. Characters and main points were identified. The students were able to give more than a description of events based on the illustrations. Just as the scores improved on the retell, so did most of the subtests of the other assessments on language.

The teachers in the study provided input on the Story Talk instrument. Some teachers stated that it was time consuming to learn how to assess the oral narratives and it was time consuming to administer the assessment of each student while the class participated as an audience. A teacher stated that she

was not confident with scoring because there were language nuances that she was not certain how to score. Some of the teachers felt it would be best to have two individuals score and reach agreement. The test took 20-30 minutes to administer. On the other hand, they had two teachers that indicated the following:

> Both teachers remarked that it has provided some really useful insights into the children's developing oral language abilities. The reception teacher, for example, remarked that one boy's retelling had changed entirely her expectations of what he was actually able to do, which was at odds with what he did in the classroom. The assessment gave the children and an opportunity to say more than they would normally be able to say in a group of whole-class situation.
> (Riley & Burrell, 2007, p. 192)

A cross-sectional study (Miller et al., 2006) that used story retell with Spanish-English bilingual children required children to engage in story retell with a wordless picture book. The researchers compared the vocabulary and narrative structure used in each language in order to compare language proficiency between both languages. The study had 1, 531 Hispanic/Latino Spanish-speaking ELL participants, from kindergarten through third grade in TBE programs, and attempted to ascertain which features of oral language were associated with reading proficiency in each language. The study's research questions focused on whether oral measures scores in Spanish predicted reading scores in Spanish and whether English oral measure scores predicted scores in English reading. In the study, the researchers examined whether there was a cross-sectional impact in terms of the oral language scores in one language (English/Spanish) predicting scores in the other language

(Spanish/English). The language measures they used collectively provided information on language performance, syntax, vocabulary diversity, general proficiency, and narrative structure. The reading measure used was the English and Spanish Woodcock Passage Comprehension from the Woodcock Language Proficiency Battery – Revised: English and Spanish (Woodcock, 1991a). In the study they found that “the oral language measures accounted for a significant amount of variance for both reading measures, with grade controlled, in both languages” (Miller et al., 2006, p.39). The researchers found that the oral language skills in one language accounted for significant variance in reading scores in the other language.

There were some limitations to that study. First the researchers had a problem with the assessment of the advanced students, because they were wanting to read the print rather than using the pictures to help them retell stories. Also, as the researchers noted, only longitudinal data could answer the question of language strength and change over time. There is also the issue of confounding variables that affect student performance based on the type of program they are in. The programs that the students were in were not defined sufficiently to know if the program of instruction and language treatment was comparable to other program models. It is possible that the sample was not comprehensive because as students became more fluent in English they were exited from ELL classes and this caused an over representation of lower performing ELLs, according to the researchers.

In 2004, Fiestas and Peña conducted a small scale study (n= 12) with bilingual children, of comparable fluency in Spanish and English, between the ages of four and six, to investigate the effect of language on narratives produced by Spanish-English bilingual children. These students produced four story retells (narratives). Two of the narratives were produce in Spanish and the other two were in English. These narratives were produced using the prompt of a picture book and a static picture. The narrative from the books were measured in terms of complexity, total words, number C-units (independent clause plus its modifier), and mean length of C-units. The transcribed narratives were coded for story grammar and grammaticality and the Systematic Analysis of Language Transcripts (SALT) was used to analyze the narratives. Repeated measures of variance (ANOVA) was used to analyze the number of C-units, MLC-words, number of words, and grammaticality. A within subject design was used with language as the within subject variable. For grammaticality the same design was used; but with task (picture book or wordless picture book) as the other within subjects variable.

First, the researchers found that the language of the narrative, when measuring story grammar, did not influence the complexity of the narrative. The story grammar ratings for the wordless picture book was similar for Spanish (M = 5.08) and English (M = 4.75). Second, for the story elements comparison with language there was a significant main effect for narrative elements $F(6, 66) =$

10.194 $p < .011$ and there was a significant language and narrative interaction $F(6,66) = 2.440$, $p = .034$.

In the study, the researchers also found that children produced comparable narratives for the book task in Spanish and English. However, the picture task yielded varied results. With the Spanish task children tended to provide initiating events and problem-solving narratives. In the English production, the students tended to provided narratives that illustrated consequences. The researchers indicated that there could be cultural influences to cause this language effect. Because children tended to elaborate more in their native language, the researchers cautioned that language and narrative tasks should be considered when testing bilingual students.

There are some limitations to the Fiestas and Peña study. First, the study used a small sample (n=12) and a sample of convenience. The age of the participants was of a wide range from four to seven years which encompasses diverse developmental expectations and limits comparisons. Research assistants transcribed the audio-recorded verbal communication from the students. However, there was no intrarater or interrater reliability check for consistency in transcribing. Interrater reliability checks were conducted for rating the transcribed student responses; however, agreement and interrater reliability were not reported. Although, the book task provided sufficient language production for analysis, the picture task was conducive to curtailed responses, which limited discourse availability for analysis. This picture task manifested

mixed results as some students produced personal stories and scripts, so it could not be compared to the book task.

Gutiérrez-Clellen (2002) reinforced that little is known about narrative performance of bilingual children, in particular, of Spanish-speaking children that are becoming bilingual. The premise of the researcher is that narrative studies are typically assessed in bilingual students' L2 (English) and this would increase the propensity of students being deemed to have low literacy when in reality they just may have inadequate L2 proficiency to express their comprehension. Therefore, Gutiérrez-Clellen conducted a study with the use of story retell and story comprehension tasks, in English and Spanish, to assess the narrative performance of bilingual children (n=33) ages seven and eight. Five of the students were receiving English-only instruction and the remaining 28 students were receiving instruction in both languages. The participants in this study were from a larger sample of participants in a story recall and story comprehension study. The researcher elicited narrative samples from two distinct books, but comparable in complexity and length. In this study, T-units (one main clause and all its subordinate clausal and nonclausal elements) were assessed. Grammatical errors were also assessed. The story recalls were analyzed using SALT. The story comprehension questions were scored using a protocol developed for the study. T-tests were used for the analysis and a significant differences existed $t(32) = 4.30$, $P < .001$ for English narratives in comparison to Spanish . The effect size of this difference was $d = 0.73$. The second t-test

compared the responses to the comprehension questions across both languages. The participants seemed to manifest greater, $t(32) = 4.28$, $p < .0002$ with $d=0.72$, English story comprehension than Spanish story comprehension. There was also greater variability in the Spanish scores than in the English scores. It was interesting that the children that performed better in one of the two languages, still scored within one standard deviation from the mean of these students in their weaker language. According to the researcher, “the data underscore the notion of bilinguals as a continuum of proficiencies…narrative assessment tasks in L1 and L2, which appear comparable, may not pose similar processing demands on a bilingual speaker” (Gutiérrez-Clellen, 2002, p. 192).

Among the limitations of that study was the small sample size (N=33) and participants being from the same district. Also, the population was homogenous in that all the participants but one was from Mexican-American descent and the majority were U.S. born. In this study a bilingual research assistant transcribed the recordings but no details were provided in terms of intrarater reliability of those transcriptions. The stories in this study were told once and the student produced a retell immediately. The outcome for the students might be different when assessing them after allowing time to pass between the story being told and the narrative being solicited.

Finally, a story retell study ensued from Project ELLA which measured the impact of the STELLA intervention (B. J. Irby et al., 2004) and was conducted as a dissertation study by Quirós (2008). This study (n=72) was with

second-grade students in a Transitional Bilingual Education program model: 37 students were in the experimental group and 35 were in the control group. The researcher assessed the story retell of the students in English and Spanish. The variables that Quirós examined were T-units, number of words, number of sentences, vocabulary, story grammar, and end-of-story assessment. The instruments that the researcher used in this study were the Naglieri Test of Non-verbal Ability (NNAT) (Naglieri, 1997), curriculum-based measures (for vocabulary, end-of-story assessment, and retellings), and teacher's observation protocol scores (customarily used in Project ELLA). In the study the dependent variable was the total number of words for the retellings. Quirós stated that T-units were not used because there could be variations for T-units in English as opposed to Spanish and they should not be compared because English will inherently yield higher T-Units.

The study used a analysis of covariance (ANCOVA) and found that there was a statistically significant difference in the length of story retell between the experimental and control group for week one and week six. The results were as follows in week one: T-units, F (1, 66) = 35.737, $p < .001$, d = 1.41; number of words, F (1, 66) = 46.572, $p < .001$, d = 1.62); and the number of sentences, F (1, 66) = 31.828, $p < .001$, d = 1.37 after controlling for non-verbal ability. In week 6 the results were as follows: T-units, F = 47.293, p < .001, d = 1.68; number of words, F = 69.346, p < .001, d = 2.03; and number of sentences (F = 23.18, p < .001, d = 1.19. Apparently, one could attribute a positive contribution

from the structured story reading on student performance. The researcher also looked at the same participants but tested to see if there was an impact on Spanish oral development, as measured by Spanish retelling. In week 1, the results revealed a statistically non-significant differences between the TBE – E and TBE-C groups: T-units, a $F(1,66) = .742$, $p = .392$ and number of sentences $F(1,66) = .386$, $p = .536$. In terms of words produced there was a statistically significant results $F(1,66) = 11.595$, $p < .001$, $d = .81$. For week 6 the results were as follows: T-units $F(1,66) = 42.357$, $p < .001$, $d = 1.58$; number of words $F(1,66) = 59.627$, $p < .001$, $d = 1.89$; and number of sentences $F(1,66) = 66.537$, $p < .001$, $d = 2.00$. For vocabulary a 20 question multiple-choice test was administered and the differences were statistically significant: $F(1,65) = 51.58$, $p < .001$, $d = 1.77$, again demonstrating greater gains for the students receiving the STEALLA intervention. For listening comprehension story elements on questions and retell the TBE-E group outperformed the TBE-T group and the results were as follows: For week 1 $F(1,66) = 72.556$, $p < .001$, $d = 2.02$ for week 6. Furthermore, the curriculum-based assessment for comprehension and vocabulary revealed a statistically significant differences between the experimental, $F(1, 66) = 32.660$, $p < .001$, $d = 1.32$ and control $F(1, 66) = 29.685$, $p < .001$, $d = 1.27$ groups.

Vocabulary

Without some knowledge of vocabulary, neither language comprehension nor language production would be feasible (Anglin, 1993). The STELLA

curriculum developers in Project ELLA agreed that “while much is known about the importance of vocabulary to success in reading, there is little research on the best methods or combinations of methods of vocabulary instruction and the measurement of vocabulary growth and its relation to instruction methods” (National Institute of Child Health and Human Development, 2000b, p.17). Therefore, STELLA in Kindergarten had as one foci oral language vocabulary development, specifically, vocabulary instruction, development, and measurement for ELLs.

Vocabulary and Reading Comprehension

Several researchers have found that vocabulary knowledge was a significant correlate of reading comprehension (Roth, Speece, & Cooper, 2002; Snow, Burns, & Griffin, 1998) and that systematic, intensive, purposeful, and effective instruction in vocabulary impinged reading comprehension (Beck & McKeown, 1987; Stahl & Fairbanks, 1986). The National Reading Panel (National Institute of Child Health and Human Development, 2000, p.16) reported on the importance of oral vocabulary and print vocabulary by affirming that “…the larger the reader’s vocabulary (either oral or print), the easier it is [for the reader] to make sense of the text” (p.16).

Vocabulary Indirect and Direct Instruction

Nagy and Herman (1985) proposed an incidental learning hypothesis which was based on research on native language learning. According to the hypothesis, most words are learned gradually through repeated exposure to

words overtime in various discourse contexts. Therefore, they began to advocate the practice of extensive reading to significantly increase vocabulary acquisition in L1 (Nagy & Herman, 1987). For L2 vocabulary acquisition, Krashen (1989) concurred with the importance of indirect vocabulary acquisition through reading. He claimed that vocabulary was acquired through comprehensible input: Input Hypothesis theory. Elley (1991) presented the results of nine studies that dealt with children acquiring vocabulary through high interest reading and found that children had rapid gains in reading and listening comprehension, and these gains remained stable overtime. Elley concluded that these studies provided support for whole-language approaches and Krashen's Input Hypothesis.

Not all researchers agreed that indirect vocabulary instruction was as effective as proclaimed by Krashen. Coady related that research that positively supported Krashen's claim was limited (1997). Furthermore, Ellis (1994, pp. 13-15) posited that it was not comprehensible input that is needed to enhance instruction, but actually “comprehended input.” Mason, Stahl, and Herman (2003) held the view that direct instruction was important and they recommended that vocabulary instruction (a) include information on definitions and context of words, (b) actively engage children in the learning process, (c) provide multiple exposures to meaning word information.

It is not clear if readiness for productive use can be reached by receptive exposure, such as with large quantities of reading or listening, or whether there

must be forced output: learners being made to speak or write (Swain, 1985). Furthermore, when the goal of instruction is for the student to produce language then there must be productive learning, thus, further limiting the comprehensible input hypothesis (Swain).

Word Knowledge

Determining what constitutes word knowledge is an initial step in the study of vocabulary acquisition, development, instruction, and assessment. Beck, McKeown, and Kucan (2002) provided two dimensions to word knowledge. One dimension was denominated “word ownership.” To have word ownership the student must have demonstrated knowledge of the words and appropriate use of the words. The other dimension was “word awareness.” At this level the student began to take notice of words in a general way . For example, a student began to notice word families and word associations when he or she encountered a new word. Bear and Helman (2004) proposed the following example to distinguish between word knowledge and word ownership: “with students in the intermediate grades, they have been around enough to be exposed to much print so they can recognize words, read them aloud, and spell them but not necessarily ‘know them or own them’” (p. 154).

“Words are not isolated units of language, but fit into many interlocking systems and levels. Because of this, there are many things to know about any particular word and there are many degrees of knowing” (Nation, 2001, p.23). Stahl (as cited in Stahl & Fairbanks, 1986) provided these three levels to

describe the depth of word processing: (a) association, which was to learn the form of a word and form a meaning connection; (b) comprehension, which was to recall the meaning of a previously met item; and (c) generation, which was to produce a novel response to an item such as restating a definition in different words or making an original sentence. These levels resembled what Nation deemed to be evidence that a word was known: noticing a word, retrieving a word, and using a word generatively (Nation, 2001, p.75). In using a word generatively the learner produces a word in a new sentence context and/or the leaner produces associations, causal link, etc. (Nation, 2001). The participants in this study were asked to use target vocabulary words, generatively, and these utterances were analyzed using the S4.

Assessment Perspectives

Read (2000) outlined two contrasting perspectives in vocabulary assessment. One perspective was focused on whether the learner knew the meaning and usage of a set of words that were taken as independent semantic units. The other perspective was grounded in the notion that words should be assessed within the realms of the context of the language-use text.

Assessment Dimensions

Read also provided three dimensions for vocabulary assessment: discrete or embedded, selective or comprehensive, context dependent or context independent. These dimensions focus on the construct that is being measured. In a discrete test, vocabulary knowledge is a distinct construct,

separated from other components of language competence. Most vocabulary tests are designed on the assumption that it is meaningful to treat words as independent constructs. In contrast, embedded vocabulary measures are those that contribute to the assessment of a larger construct. Tests that measure embedded vocabulary ask about the meanings of certain words. Even if the words are presented as part of a reading comprehension exercise, the questions ask the examinee to determine the meaning of a word based on the context. Here, the score for the vocabulary questions is just a part of the whole comprehension measure. In addition, a test can have a large amount of context, like a long reading passage, and if all the questions are focused on conveying the meaning of a word (without needing to rely on the context), then that measure becomes discrete instead of embedded. Thus, to determine whether a particular vocabulary test is discrete or embedded, one needs to consider the purpose and the way the results are interpreted.

The second dimension distinguished between selective and comprehensive, and it takes into account the range of vocabulary to be included in the assessment. An example of selective is the conventional vocabulary test that is based on a set of target words selected by the test-writer, in which the test-taker is assessed according to how well she or he demonstrates knowledge of these words. Comprehensive measures take into account all the vocabulary content of a spoken or written text. For example, an interview where particular words are not assessments, but instead mark quality or overall vocabulary, is

judged as an example of a comprehensive measure. Another example is measuring the number of *sophisticated* or *low-frequency* words used by the examinee. Readability formulas are also an example of comprehensive-embedded measures (Read, 2000).

The third dimension considers whether the language is context independent or context dependent. Contextualization is more than where vocabulary is presented. “The key question is to what extent the test takers are being assessed on their ability to engage with the context provided in the test” (Read, 2000, p. 11). Can the test taker give appropriate responses as if the words were in isolation, or is the text needed? An example would be when the answer choices on a multiple-choice test are all appropriate definitions or synonyms of the target word, and the examinee is then required to look at the context to decide which meaning is applicable.

In summary, before designing an intervention, commencing research, or approaching assessment distinctions, decisions must be made. The decisions Read recommended were as follows: (a) deciding whether to measure receptive and/or expressive language, (b) choosing what words to teach and test, (c) asking does/should the instrument test these words independently, or are there other language components that factor in to enhance or detract from performance, (d) settling on what perspectives of Vocabulary Assessment will be addressed, such as independent semantic units or in context language use as defined by Read, and (e) selecting which dimension of vocabulary will be

assessed, such as discrete or embedded, selective or comprehensive, and context independent or context dependent.

Choosing Words to Teach

Some may think that measuring advanced vocabulary or word production is the solution, but actually simple words also are considered advanced (Read, 2000). This makes it difficult to determine what types of words to include in an intervention and assessment. A researcher must be able to explicate and justify the words that are targeted in an intervention and assessment. Whether words are basic, or highly academic words, or fall anywhere along the continuum between basic and highly academic words, they are all valid targeted vocabulary.

Vocabulary Word Levels

The vocabulary instruction in STELLA was influenced by the Beck and McKeown (Beck et al., 2002) three tier categorization of word difficulty. Tier I includes words that have high frequency and that a student would be expected to know the word based on encountering the word on a regular basis. Tier II words are high frequency words, but they are not basic words. Tier III words are not encountered with frequency. These words are usually related to content. For STELLA and when working with ELLs, it is important to note that Tier I words should be part of instruction because second language learners may not know words at the Tier I level (Irby et al., 2008).

Vocabulary and ELLs

Reading comprehension (in L1 and L2) is affected by the reader's background knowledge and use of reading strategies such as: Prediction, deciphering unknown words in context, making inferences, recognizing text types and text structure, and identifying the main idea. Yet, it has been consistently demonstrated that reading comprehension is strongly related to vocabulary knowledge, more strongly than the other components of reading (Laufer, 1997).

Despite that vocabulary has been established to be of paramount importance to the language learner, teaching and learning of vocabulary have been undervalued in the field of second language acquisition (SLA) (Zimmerman, 1997). Zimmerman stated that students cannot be expected to learn by themselves; second language learners need to be provided systematic vocabulary instruction. Particularly with second language learners, "they cannot be expected to 'pick up' substantial or specific vocabulary knowledge through reading exposure without guidance" (Paribakht & Wesche, 1997, p.177). Coady (1997, p. 229) elaborated on the concern of beginner second language learners in light of the empirically based and supportive evidence in incidental acquisition. Coady stated that "beginner learners are in a paradox, in a beginner's paradox, because how can they be expected to learn sufficient vocabulary through extensive reading when they do not know a sufficient amount of words to read well?" (p. 229). Coady added that a pragmatic

approach would dictate that for learning L2, the focus should be on words. However, most contemporary academic approaches to language learning placed minimal importance on vocabulary learning and appeared to assume that somehow words would be learned as a by-product of the other language activities (Zimmerman, 1997). Laufer (1997) provided an estimate of the number of word families that a good L1 reader needed to know in L2 in order to read well; that estimate was 3,000 word families or about 5,000 lexical items.

Carlo et al conducted a quasi-experimental study, Vocabulary Improvement Program, with fifth-grade ELL students *(n=142)* and English-only students *(n=112).* These students were in 16 classrooms (10 experimental and 6 control) in three distinct sites. The students in the control classrooms received instruction as part of the normal school curriculum. The students in the experimental settings received vocabulary instruction over the course of 15 weeks, in which 10 to 12 target words were introduced and taught four days per week for 30 to 45 minutes. Three times during the study (at each 5th week mark) a comprehensive review of words was conducted. The intervention was designed around the topic of immigration and included readings from newspaper articles, diaries, and immigration documented accounts. The intervention included detailed lesson plans and *quasi-scripted* lesson guides. The specifics of the program included previewing an assignment in the student's native language on the first day of the lesson cycle. On the second day, students read in English, target vocabulary was introduced, and large group discussion took

place in regards to those target vocabulary words. On the third day, students worked in small groups and completed cloze activities (filling in the blanks on sentences). On the fourth day, students completed word association, synonym/antonym, and semantic feature analysis activities. On the final day, the students partook in sundry intervention activities that promoted word analysis skills, rather than the learning specific target words.

The measures that were used in the Carlo et al study were the Peabody Picture Vocabulary Test Revised (PPVT-R) (Dunn, Dunn, Robertson, & Eisenberg, 1981), a Ploysemy production measure (in which students produced as many sentences as the possibly could while conveying the different meaning of words), a Reading Comprehension multiple-choice cloze passages measure, Word Mastery measure in which students selected the definition that best corresponded to a word from four answer choices for each of 36 target words, Word Association task (Verhallen & Schoonen, 1993) in which target words were matched with other words that were closely connected or associated, Morphology was tested using a modified version of Extract-the-Base (Carlisle, 1988) task.

A multivariate analysis of variance was used in the study. When examining time x condition the results for reading were F (1, 213) = 17.84, p<.001 and η^2=.08. The results for Mastery were significant F (1, 218) = 113.28, p<.001 and η^2=.34. The results for Word Association were significant F (1, 217)= 11.24, p<.01 and η^2=.05 and also significant for Polysemy. For

morphology the results were not significant (p>.05). When just examining time of test, pre and post, for all the above measures Word Association, Polysemy, and Cloze were not statistically significant (p>.05). The only measures that were statistically significant were Mastery F (1, 218) = 7.64, p<.01 and η^2 =.03 and Morphology F (1, 217)= 11.46, p<.01 and η^2 =.05.

Carlo et al stated that a limitation of their study was that researchers Shanahan, Kamil, and Tobin (1982) questioned the valid use of cloze activities to measure reading comprehension. In the study there was no indication that the correlation among dependent variables was examined and reported. In MANOVA power can be affected by the correlation between the dependent variables and the effect size (Cole, Arvey, & Salas, 1994). Also, there was no theoretical or empirical support for lumping the dependent variables in this study as suggested by Field (2005). Any of these two factors can influence the accuracy of the MANOVA test statistics.

In another study, Loftus (2008) studied vocabulary instruction with kindergarten students (*n = 43*) (from a school where 70.7% of the students are Hispanic). The participants in the study were deemed at risk for language learning and with them Loftus compared the difference between a Tier 2 vocabulary intervention and research-based Tier 1 vocabulary instruction. The interventions in this study were based on the Response to Intervention (RTI) model (Mellard, Byrd, Johnson, Tollefson, & Boesche, 2004) which is used for early identification of students who may be at-risk for learning difficulties. Tier 1

instruction is researched based general classroom instruction. As students are identified at-risk, they receive Tier 2 instruction if they do not respond exclusively to Tier I instruction. In proportion to a student's propensity to be at-risk, the Tier level of instruction can increase up to level 4 (Marston, 2005).

In the Loftus study, all the students received research-based, Tier I instruction. The students that scored less than a standard sore of 92 on the Peabody Picture Vocabulary Test – III (PPVT-III) (Dunn, Dunn, Williams, & Wang, 1997) were considered at-risk and received small group, Tier 2 vocabulary instruction in addition to the Tier I instruction. There were 20 students that had scores below the cutoff mark. In Tier 1 instruction, in which all participants were instructed *(n=43)* students listened to two storybooks read to them, twice, during a two-week period. Each book contained four target words for a total of eight target words in a two-week period. The additional Tier 2 intervention required the participants to work in small groups (three to four students). These students met with an intervention specialist for 30 minutes per day. During the session with the interventionist, two of the words from each book were taught via vocabulary activities the other two words were not.

Word knowledge was assessed using the researcher's measures of receptive and expressive vocabulary. These measures were administered after the first week of the intervention and again seven weeks later. The measure for Specific Word Knowledge was similar to the one in my study, in that it also examined word knowledge along a continuum. Loftus included a Word

Recognition measure in which students verified if they recognized a word. Nonsense words were included in this measure as distracters. The researcher reports that the Cronbach's alpha internal consistency coefficient for this sample was .60. The Target Word Picture Vocabulary measure required students to identify a picture that represented the target word that they were given. The Cronbach's alpha internal consistency for this was .55. To test whether students could respond to a question that contained a target word, the researcher created the Context Questions Measure which had a Cronbach's alpha internal consistency coefficient of .62. The Expressive Definition measure in the Loftus study resembles the S4, in my study, in that it rates student expressive responses by assessing points based on depth of knowledge. In this study, zero points were given for a response that was unrelated. One point was given for a response that was related to the target word. And two points were given for responses that were complete.

Results were analyzed using repeated measures ANOVAs. The within subject factors were (a) classroom-based Tier 1 instruction versus classroom-based Tier 1 instruction plus additional Tier 2 intervention, and (b) posttest versus delayed posttests. In the Word Recognition Measure the students that were receiving Tier 2 instruction scored significantly higher than those in only Tier 1 instruction $F(1,19) = 8.30$, $p=.01$ and $d = .63$. With the Target Word Picture measure there was a slight significant difference between both groups, $F= (1,19) = 2.19$, $p=.16$ and $d= .66$. In terms of the Context Questions Measure

it was statistically significant F (1,19) = 4.96, p = .04 and *d* = .42. With the Expressive Definition Measure there was a significant difference F= (1,19) = 6382, p = .02 and *d* = .69.

Students in the Tier 2 intervention made greater gains in word knowledge than those who received the traditional classroom-based instruction. The findings in this study confirm that direct instruction with young learners is important, particularly for at-risk students. However, the sample in the Loftus study was not random and it was small; thus, the findings could not be generalized to other populations. The small sample size also limited the researcher's ability to look at between-subject comparisons. Finally, the scores on the researcher created measures were not validated via common principles of psychometric properties (aside from internal consistency).

Oral Proficiency

Bialystock (1991) claimed that the way researchers define the construct of language proficiency should determine how it is viewed, measured, and taught.. Not establishing a clear construct with parameters hinders the progress of research. In a research synthesis on oral language, Saunders and O'Brien (2006) conducted a search for studies on oral language, and they found that studies on oral proficiency were one-fourth of the studies recovered for literacy in general. They found 150 studies on oral language development, fewer than two-thirds of those studies reported oral outcomes, and fewer than one-third met the criteria of reporting language outcomes and being seen as relevant and

methodologically sound. They found it difficult to create a synthesis and to generalize the studies of oral language proficiency research because some studies reported general oral proficiency, while other studies reported discrete elements of oral language proficiency, and others measured language use and/or language choice (2006). Saunders and O'Brien (2006) ascertained that the Snow et al. study "is one, if not the only, attempt in this corpus to operationalize and examine empirically the nature of oral language use for academic purposes" (p. 17).

In a recent study (Tong, Lara-Alecio et al., 2008b), which utilized data from project ELLA on oral language proficiency. In the Tong study with Hispanic ELLs students (*n=534*) researchers examined growth trajectories and rates on academic English oracy using latent growth modeling. The researchers compared student performance under two program models: experimental TBE and SEI and control TBE and SEI. The students in the experimental and control TBE and SEI all showed statistically significant ($p<.05$) linear growth from kinder to first-grade. The students in the experimental TBE and SEI developed at a faster rate than their counterparts that were just receiving typical instruction ($p<.05$, effect sizes >0.46). In this study it became apparent that first language instruction did not hinder second language instruction and that enhanced instruction in TBE and SEI programs can accelerate oral English acquisition and help alleviate the disadvantage of students with low English proficiency. The measures that were used in this study were the Woodcock Language

Proficiency Battery – Revised (WLPB-R) picture vocabulary and listening comprehension subtests.

Oral Proficiency and Reading

Loban (1976) conducted a longitudinal study on the language development of children from kindergarten to grade twelve. In his study, he found that children with advanced language ability in the early years, were flexible with *movables* of language in subsequent years. *Movables of* language are parts of sentences that can occur in several different places. An example of movable language is the following: (a) *Susan opened the door with great care;* (b) *With great care, Susan opened the door;* or (c) *Susan, with great care, opened the door.* Children with this skill were found to have greater reading achievement from year to year. Some researchers (Miller et al., 2006; Snow, 1983) have found that L2 communicative competence established the foundation for subsequent literacy development, and if measured at a cognitive academic language proficiency (CALPS) level there was a greater association between oral proficiency and reading achievement (Riches & Genesee, 2006).

Difficulties in oral language development seem to indicate a propensity for difficulties with reading (Biemiller, 2003; Catts, Fey, Tomblin, & Zhang, 2002). Vocabulary, syntax, and idiomatic comprehension are some measures of oral language that have been attributed with predicting reading achievement (Durgunoglu, Nagy, & Hancin-Bhatt, 1993; Lindsey, Manis, & Bailey, 2003; Manis, Lindsey, & Bailey, 2004; Proctor, Carlo, August, & Snow, 2005). Even

when the target reading is not in English, a study (Vaughn et al., 2006) showed that a daily, 50-minute, Spanish intervention with an oral language and reading focus showed that students with Spanish skills that were 1.5 *SD* below expected levels could advance to near average-levels, .08 *SD*. However, in this study oral proficiency measured both expressive and receptive language, and students manifested greater gains with receptive language.

Discrete Elements of Oral Proficiency

Developing oral proficiency in English involves the acquisition of vocabulary, control over grammar, and understanding of subtle semantics (Saunders & O'Brien, 2006). Oral interactions have also been an integral part of acquiring English oral proficiency, these interactions included: exchanging greetings, initiating and sustaining conversations, negotiating collaborative tasks, giving and receiving directions, and telling and listening to stories (Saunders & O'Brien, p. 14).

Vocabulary. Mason et al (2003) recommended that children learning vocabulary should be viewed in the context of overall development in literacy because learning is integrated and not disjointed, and oral language development has an inextricable link to literacy development (Pinnell & Jaggar, 2003). "Deep similarities exist between word learning and other aspects of language development …words are learned through abilities that exist for other purposes. These include an ability to infer the intentions of others, an ability to

acquire concepts, an appreciation of syntactic structure and certain general learning and memory abilities" (Bloom, 2000, p. 10).

According to Read (2000), productive vocabulary is the set of words that an individual can use when writing or speaking. They are words that are well-known, familiar, and used frequently. Conversely, receptive or recognition vocabulary is that set of words for which an individual can assign meanings when listening or reading. These are words that are often less well known to students and less frequent in use. Typically, these are also words that individuals do not spontaneously use. However, when individuals encounter these words they do recognize them, even if imperfectly (Read, 2000).

The capacity of ELLs to define words has been a measure of proficiency. Vocabulary development studies have shown that the capacity to define words and the formality in defining words increases with proficiency (Saunders & O'Brien, 2006). At lower levels of proficiency, ELLs have tended to define words in terms of associations. As students increased in proficiency, the type and quality of the definitions that they provided evolved, as evident in a study by Snow et al (1987). In a study, Snow et al (1987) asked 137 second- to fifth-grade students to provide an oral definition to common words. The researchers coded the student responses as either formal or informal and rated the responses on quality. The definitions that rated highly were those that were formal, had sophisticated vocabulary, and elaborate syntax. These elements were deemed as indicators of academic language because they were

decontextualized. Similar to children's communicative and conversational skills, vocabulary development appears to be protracted, "becoming more impressive after the child had entered school than before"(Anglin, 1993, p.2).

When evaluating vocabulary and oral proficiency knowledge and competence it is evident, as Read (2000) has indicated, that often the words tested are predominately based on written vocabulary. The distinct characteristics of spoken vocabulary have not received much attention, by comparison. Much of the research on vocabulary has been undertaken by reading researchers, who obviously focus on words in written texts. There is no equivalent research tradition examining the vocabulary of spoken language, especially in informal settings (Read, 2000).

Grammar. The Oxford Dictionary defines grammar as "the whole system and structure of language or languages in general, usually taken as consisting of syntax and morphology (including inflections) and sometimes also phonology and semantics" (Lindberg, 2000, p. 580). Also, Close (1982, p. 13) stated that "English grammar is chiefly a system of syntax that decides the order and patterns in which words are arranged in sentences."

Hawkins (2001) related there were similarities in the grammar-building of first language and grammar-building of second language and that the principles of Universal Grammar could be applied to the study of second language acquisition. "An important part of learning a second language is learning how words fit together to form phrases, and how phrases fit together to form

sentences. The combinatorial properties of words and phrases are known as the syntax of a language"(Hawkins, 2001, p.1). Sentences are constructed with the syntactic properties of a given language and are grammatically correct, grammatical. If the produced sentence violates the correct construction then it is grammatically incorrect, ungrammatical (2001).

Dulay and Burt (1973) studied how often Spanish-speaking ELL children used eight grammatical morphemes in an appropriate way. They found the plural *s* to be the easiest morpheme for the learners (Girls go). Progressive *ing* was the next easiest in present tense used in the word *going*. Next the copula forms of *be* meaning the use of *be* as a main verb in a sentence (John is happy) as opposed to its use as an auxiliary for another verb (John is going.). The auxiliary from of *be* with *ing*, such as with *Girls are going* is another example." Fifth in difficulty was the tense of definite and indefinite articles *the* and *a* to produce such sentences as *The girls go* or *A girl go*. Sixth, was the use of the irregular past tense. Those words that did not end with /d/ were still pronounced as such (with /d/, /t/ or id). The seventh in difficulty was the use of third person used with verbs, as in *the girl goes*. Finally, the last area of difficulty was with possessive *s* and with the *s* ending used with nouns to show possession, as in *The girl's book*. The first language background does not make a difference in the progression of difficulty for children; nor did language background influence this process for adults. In my study, these types of idiosyncratic patterns of

second language acquisition were considered in the development of the S4 scale descriptors.

Syntax. Syntax is the study of how morphemes combine to form sentences (Piper, 2003). “An understanding of syntax, for instance, allows us to produce and understand a potential infinity of new sentences” (Bloom, 2000).

“Children’s syntactic performance is based on the rules that children use to combine words into phrases and sentences. Just as with assessing phonology and morphology, assessing children’s syntactic knowledge requires collecting examples of syntax in use” (Duchan, 2004, p.55). Skinner (1957) proclaimed that children acquired language through classical and operant conditioning which involved the process of children making sounds and eventually imitating parents, all the while receiving either positive or negative reinforcement which is what facilitated language and grammar learning and production. In opposition to this, behaviorist view of language differed and was similar to the view of Noam Chomsky. Chomsky founded the branch of cognitive psychology and believed that children were predisposed to learn language via the use of mental slates that contained the necessary, genetically inscribed knowledge needed for language (Adamson, 2005). In 1957, Chomsky published *Syntactic Structures* (Chomsky, 1957), and in this book and in his subsequent research, he proposed that mental processes could be studied, beyond the realms of merely looking at behavior, and that language provided a window into the mind. Chomsky believed that children were born with specific linguistic

knowledge and a predisposition for it and he called this universal grammar (Adamson, 2005). Chomsky's goal was to develop a system of grammatical analysis: generative grammar which is what the child must learn in order to resemble the grammatical production of the adults in the child's culture. Some concerns with Chomsky's theories have been that they leave out the human interactional element, how grammatical forms are used to accomplish human goals. Therefore, the theoretical framework, as described in Chapter I of my study, encompasses the human interactional and communicative element of these theories under communicative competence. Communicative competence was originally attributed to the work of Canale and Swain (1980) and later expanded by Bachman (1990).

By the time that L1 English-speaking children are 6-years-old, they have mastered the basic syntactic structures of English (Piper, 2003). Syntactic development can be viewed through the perspective of global clause structure of sentences and development within their two major constituents, the noun phrases and the verb phrase (Piper, 2003).

> Clause structure refers to the patterns that children use in constructing their sentences – Subject-Verb-Object, Subject-Verb-Object-Complement, etc. Research has shown that although children are able to understand and produce a wide range of clause patterns by age six, they typically produce only a few. The predominant ones are transitive SVO sentences with or without a sentence adverbial. They also produce a number of intransitive sentences, with and without adverbs, but other structures, such as sentences with two objects (Mathew gave Spook the food or Grammy gave that book to me) are less common. Adverbs play an important role in young children's clause structure, and the use of adverbs has been widely studied. The kinds of adverbs children use appears to be established by age six and remains unchanged

> throughout the elementary school years. The proportion of adverbs defining place, time, manner, and cause or condition is largely predictable from the order in which they were acquired – place and time first at about age two and then by age six manner adverbials, and then those expressing cause or condition.(Piper, 2003, p. 114)

Semantics. Semantics can de defined as "the study of the relationship between linguistic signs and the real world [and] it is impossible to study ANY aspect of language without considering meaning…" (Piper, 2003, p.51). Some linguists believe that semantics and syntax must be viewed and studied in an integrated manner (Piper, 2003). Conceptualizing semantics and syntax as integrated is part of generative theories and case grammar. Both of these theories hold that syntax is determined by semantics. Piper further elaborated that the study of semantics is difficult because semantics cannot be precisely formalized. Semantics reside in words, sentences, and larger units of discourse, singly or simultaneously (Piper, 2003).

Oral Proficiency Measures for ELLs

Cummins (1981) explained that when assessing oral proficiency, solely assessing natural communication is not sufficient because natural communication occurs at a Basic Interpersonal Skills Level (BICS) and literacy skills occur at a Cognitive Academic Language Proficiency Level (CALPS). Therefore, he contended that typical measures such as the basic Inventory of Natural Language (BINL) and the Bilingual Syntax Measure (BSM) should not be used to make program placement and exit decisions. Schrank, Flethcer, and Alvarado (1996) in a study examined the BICS and CALPS comparison of oral

proficiency measures . These researchers examined the validity of the Idea Oral Language Proficiency Test (IPT – 1) (Ballard, Tighe, & Dalton, 1980), the Language Assessment Scales (LAS) (De Avila & Duncan, 1991), and the Woodcock Language Proficiency – Revised (WLPB-R) (Woodcock, 1991a). These tests were chosen because they are often used for program placement. The participants in the study were 77 kindergarten bilingual students and 199 second-grade bilingual students. The L1 for these children was Spanish and the L2 was English. In this study the researchers compared the IPT, LAS, and WLPB-R to the Language Rating Scale (LRS) which the researchers obtained from Houston Independent School District. The LAS is a likert-scale (1-5) that looks at language ability in terms of sentence structure, vocabulary ability, recalling words, telling stories, idea formation, and speech. This instrument was rated by teachers and was considered to adhere to CALPS principles. With the kindergarten sample, they found the higher correlation of (.80) to be between WLPB-R and LRS. The LRS correlation with the pre LAS was .74. Kindergarten students were not administered the IDEA. With the second-grade participants, the researchers found that the correlation with LRS was .80 with WLPB-R, .76 with LAS, and .68 with IDEA. The researchers cautioned against the use of BICS-type measures for high stakes decisions making, such as program placement and program exit because measures that evaluate CALPS provide a better picture than those that measure BICS (Schrank et al., 1996). The bellwether of oral proficiency is the production of oral language in a

communicative context where multiple levels of language use can be observed and where production can be measured in terms of words, sentences and narrative structure particularly when dealing with bilingual students (Dockrell & Messer, 2004; Miller et al., 2006).

Curriculum-based Assessment

In my study, as in the corpus of relevant literature, curriculum-based assessment was interchanged with the following terms: curriculum-based assessment (CBA), classroom assessment, formative assessment, and alternative assessment. In addition, assessment was interchanged in this review, as in the literature, between two words: assessment and measure.

Stiggins (1999, p.193) provided a comment that was a perfect segue into the importance of curriculum-based assessment and justified the need for assessments such as the STELLA Vocabulary Oral Proficiency Measure. The comment was as follows:

> We [*sic*] have centered so heavily on the development of ever-more-sophisticated psychometrics and test development tactics for our high stakes tests that we [*sic*] have almost completely ignored the other 99.9% of the [formative] assessments that happen in a student's life. These are the assessments developed and used by their teachers in the classroom. If we seek excellence in education, then the time has come to invest whatever it takes to assure that every teacher is gathering dependable information about student learning, day-to-day and week-to-week, and knows how to use it to benefit students. This action must be central to all future school improvement efforts, because if assessment is not working effectively in our classrooms everyday, then assessment at all other levels (district, state, national or international) represents a compete waste of time and money.

Purpose of Tests

Tests are intended to function as formative, summative, or predictive measures.

> Formative testing refers to assessment on an ongoing basis, as part of the learning process in the classroom. Summative testing is aimed at examining the extent to which the student has acquired the material covered in the classroom. Predictive testing provides information about the probable future performance of the test taker, in college or in other contexts. (Shohamy, 2001, pp. 32-33).

The terms achievement and proficiency describe additional distinctions that can be made in the context of language testing. Achievement refers to the mastery of the language learned in specific course of study, while proficiency seems to measure the language competence that the student will bring to real life in a specific, future, well-defined context (ibid).

Purpose of CBA

Three purposes for assessing children with CBA were established. The first was to understand the development of a given child. It is useful for teachers to be able to identify the emerging areas of knowledge and development. Second, was to assess the progress that a child is or is not making with a given intervention. The final purpose of assessment was to identify students at-risk for academic failure. The STELLA Vocabulary Fluency Measure (B. Irby, Lara-Alecio, R., Quiros, A. M., Mathes, P. G., & Rodriguez, L., 2004) and the Semantic Scoring System (S4) (Walichowski et al., 2007) are instruments that are curriculum-based and lend themselves to fulfill the purposes of assessment.

Assessment should facilitate formative, summative, diagnostic (Bloom, Madus, & Hastings, 1981) and preliminary (Oosterhof, 2001) evaluations which are also uses for the STELLA Vocabulary Fluency Measure. In general, when looking at language evaluation for second language learners the purpose should be to make instructional decisions: “to make choices that will improve second language teaching and enhance second language learning” (Genesee & Upshur, 1996, p.4).

Curriculum-based measures, when used to monitor student progress on a given skill, have required repeated administrations of equivalent measures, and many times it has been difficult to create parallel tests with other formats and that has been another pragmatic advantage of curriculum based measures (Roberts et al., 2005).

Standardized and Commercial Tests

There have always been concerns with the utility of commercially available norm-referenced tests. These concerns have been documented in the school psychology literature (Deno, 1985; Rosenfeld & Shin, 1989; Shapiro, 1990). Roberts, Good, and Corcoan (2005) elaborated on the short comings of traditionally used, normed referenced tests: they are limited in their ability to inform instructional recommendations, underlying construct assessments are vague and thus might fail to address important skills, they don’t facilitate the ability to identify specific instructional needs for developing or modifying

instructional plans, and these measures can't be used repeatedly because they are expensive and time consuming.

Standardized tests are pervasive in schools despite their many problems: they don't inform instructional change in the classroom, they don't inform how students learn, what is needed to learn, or the best way to instruct (Goodman & Carey, 2004). Researchers (Bracey, 1989; Shepaz, 1991) reported that standardized and commercial measures cannot be used repeatedly, are often expensive, and time-consuming. Curriculum-based measures are less subject to these limitations. CBAs can be cost-effective, time-efficient, and instructionally focused. CBAs represent a useful tool for a profession that spends more time on assessment than on any other task. Standardized tests have not been effective in assessing higher-order thinking and problem-solving skills and they have promoted instruction that has focused on memorization of basic and isolated facts (Bracey, 1989; Shepaz, 1991).

Quality CBA

McMillan (2007) defined quality assessment as assessment that adheres to specific psychometric standards: validity and reliability, among other principles. Accordingly for teachers, the measure of quality of a test exceeds psychometric soundness and requires that the test assess what students can do based on the curriculum with the intent of informing instruction. An expanded definition of quality assessment had the following criteria: (a) clear and appropriate learning targets, (b) appropriateness of assessment methods, (c)

validity, (d) reliability, (e) fairness, (f) positive consequences, (g) alignment, and (h) practicality and efficiency (McMillan, 2007). Brown (2004, p.19) called attention to the attributes of effective tests by saying that they should be practical tests and not excessively expensive, stay within appropriate administration time constraints, be relatively easy to administer, and have a scoring/evaluation procedure that is specific and time-efficient. Tinajero and Hurley (2001) outlined three specific purposes for authentic assessment: (a) the measures need to be an integral part of instruction; (b) the measures need to consider the learning context of the individual child, whether they are working alone or with others; and (c) assessments must provide insight into the development and growth of language and academics.

Teachers and CBA

Researchers (Airasian, 1991; Shepard, 1995; Stiggins, 1999) have indicated that classroom-based assessment has potential for accurately ascertaining student knowledge and competence. However, O'Neil (1992) informed that most classroom-based assessments methods tended to be informal and teachers needed increased expertise in this type of assessment. Teachers should design instructional modifications based on assessment data in order to help students improve (Frey & Hiebert, 2003). However, this has been unusual for teachers to do because most teachers do not make inferences or interpret data for planning instructional interventions (Butler & McMunn, 2006). Teachers need to continuously evaluate the strengths and weaknesses

of their students and adjust their teaching in order to meet the language and literacy needs of the students (Fillmore & Snow, 2000), and classroom-based assessment helps teachers identify instructional needs and modify instruction (Hurley & Tinajero, 2001).

Assessing ESL and Young Children

ESL students have been considered as having a tenuous opportunity at academic achievement. These students have been labeled as *special* or *at risk* so monitoring their progress has become a focal point (Genesee & Hamayan, 1991). Classroom-based assessments benefit second language learners because this vinculum facilitates the integration of many learning dimensions as they relate to language proficiency (Hamayan, 1995). The process of assessing and evaluating young children is complex (Gullo, 1994), and it must adhere to some fundamental issues: (a) a match between a given child's stage of development and the method of assessment used, (b) the effects of the use of the assessment results on the child and (c) the relationship between assessment and curriculum. In essence, "any student identified as a slow learner, low achiever, or even as gifted and talented must have their needs met. An appropriate intervention and alternative assessment in the early childhood years will contribute to the reading success of ELLs with reading problems" (Irby et al., 2008, ¶ 3)

Considerations in Language Testing

The remainder of this chapter includes considerations in language testing, particularly those that are applicable to the Semantic and Syntactic Scoring System (S4), which is a foci of my study. “Most oral tests are designed with some specific purpose in mind” (Madsen & Jones, 1981, p.15). Despite the purpose and intent, it is important to acknowledge that poor tests can provide useless and meaningless information; therefore, a concerted effort should be made to create and use tests that adhere to APA established standards (see American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.), 1999; Walsh & Betz, 2001).

Psychometric Considerations

Validity. “Validity refers to the extent to which the test we’re using actually measures the characteristic or dimension we intend to measure” (Walsh & Betz, 2001, p.56). Cronbach (1971) made an important clarification on validity; he stated that “One does not validate a test…one validates and interpretation of data arising from a specified procedure” (p. 477). This was corroborated by the following, “Validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests” (American Educational Research Association et al., 1999, p. 9). Palmer

and Groot (American Educational Research Association et al., 1999) numerated some of the elements that have affected validity and some of these are (a) the test itself, (b) the test setting, (c) characteristics of the examiner, and (d) the inferences that have been drawn from the test. Some of the types of validity are content validity, criterion-related validity, construct validity, convergent and discriminant validity, incremental validity, face validity, and interpretative validity (American Educational Research Association et al., 1999). Oral proficiency tests require careful attention to factors that could influence validity because some tests, such as the Oral Proficiency Interview instrument, could not demonstrate validity "because it confounds abilities with elicitation procedures in its design and it provides only a single rating, which has no basis in either theory or research" (Bachman, 1988, p.149).

Reliability. "Reliability involves the extent to which we are measuring some attribute in a systematic and therefore repeatable way" (Walsh & Betz, 2001, p.47) . Accordingly, in classical test theory, reliability is measured under three assumptions: (a) each person or environment that is measured has some quantity of the quality that is being measured, in other words, a true score; (b) every observation of a quality or characteristic contains some degree of error; (c) the observed score reflects both the true score and some error (Walsh & Betz). Lyman (1978) delineated five major factors that led to error in test scores: time influence, test content, the test examiner or scorer, the testing situation/environment, and the examinee. In order to minimize the amount of

error one should create a reliable test by (a) developing detailed instructions for test administration and scoring, (b) ensuring test administration and scoring fidelity, (c) by creating an environment that is conducive to optimal performance for the examinee, as opposed to one that may be uncomfortable or full of distractions, and (d) the examinee should understand the instructions and have a desire to perform well (Walsh & Betz, 2001).

Rating Scales

The S4, which is a scale, is in essence a test of expressive vocabulary and oral language proficiency. Genesse and Upshur (1996, pp. 144-145) defined tests and measurement devices as describing "… attributes or qualities of things and individuals by assigning numbers (or scores) to them." Furthermore, Genesee and Upshur stated that the domain of language is very large and complex. In order to measure the domain of language, it needs to be broken down to skills that become a precise set of tasks that can be measured. For my study, language was identified as domain, with a set of skills, and then with tasks to be measured using a rating scale. There are many considerations to take in the development of rating scale such as rating scale types (criterion-reference and norm-referenced), issues with scoring, common rating scale problems, use of raters, and use of teachers as raters.

Criterion-referenced and Norm-referenced Rating Scales. Furthermore, in terms of oral language testing, it is recommended that criterion-referenced scoring scales should be used in place of norm-referenced scales. First,

criterion-referenced scoring scales allow for interpretations of the individual's degree of mastery of specific language abilities within a given domain of language competence (Bachman & Clark, 1987). Criterion-referenced language tests should provide feedback into (a) teaching decisions, (b) reporting to, and discussing achievement with parents, (c) identifying children needing special support and the type of curriculum support they need, (d) identifying children for accelerated learning, and (e) informing about standards in the class in terms of curriculum development through a subject (Baker, 1995).

The difference between standardized norm-referenced and criterion-referenced tests of languages is that in standardized, norm-referenced tests, children are often compared to a native speakers (Baker, 1996), and this practice is unfair and invalid (Grosjean, 1985). However, in practice, criterion-referenced tests can be used for comparisons. The advantage is that such comparisons are made with the intent to "facilitate feedback to the teacher that directly leads to action"(Grosjean, 1985, p. 28).

Scoring. Mechanical and human language proficiency test scoring can be either simultaneous or delayed (Clark, 1975). Simultaneous scoring happens as the exam is administered or immediately after. In delayed scoring the examinee is recorded for later evaluation. There are advantages and disadvantages to both. My study used the delayed scoring which is beneficial in terms of enhancing scoring reliability because (a) examinee's appearance or mannerisms do not influence the rater [more of an influence in some testing

situations than others], (b) tapes can be randomized or intermingled with other tapes, (c) delayed scoring allows for repetitive playback to resolve doubts (Clark, 1975).

Rating Scale Problems. North (2000, p. 13) described the challenge of developing rating scales as "trying to describe complex phenomena in a small number of words on the basis of incomplete theory". Brindley (1998) detailed that it is not always easy to determine what scale descriptors are meant to describe, what learners ought to be able to do at the different levels as opposed to what they, in fact, actually do. Thus, they may also reflect the developer's beliefs and assumptions about language learning (Brindley, 1998). Nevertheless, since scales express the developer's understanding of how good performances differ from weak ones, they form part of their definition of the construct assessed in the test (Brindley).

Raters. Bock and Bock (1984, p. 337) argued that "…human judgment is always fallible, as a result evaluation of communication has certain errors associated with it." Brooks (1957) observed that with scale use, rater's accuracy is hinged on the rater's ability to discriminate among the categories or levels provided in the instrument. "The use of rating scales then, appears to rest on the assumption that an observer is a good instrument of quantitative observation, that he or she is capable of some degree of precision and some degree of objectivity" (Bohn & Bohn, 1975, p. 343).

Teachers as Raters. Test scoring can also affect score use and interpretation because they are scored by humans raters (Chalhoub-Deville, 1996). Trained teachers are usually asked to assess learner's L2 ability. Teacher training can influence teacher's assessment and cause them to have different judgments than non-teaching native raters (Engber, 1987; Shohamy, Gordon, & Kraemer, 1992); some researchers found this to be specifically the case for L2 oral testing assessments of ability (Galloway, 1980; Hadden, 1991) However, in the Chalhoub-Deville (1996) study, her findings did not differentiate between teacher and non-teacher rater. She attributed the inconsistency to possibly the language that she studied. She studied two languages in Arabic. Arabic has a diglossic situation because both languages co-exist, and one is used for formal communication and the other for quotidian communication, but not readily understood by all Arabs (Chalhoub-Deville). So the distinctions that the raters were making in this study were more overt, perhaps attributable to the consistency between teacher raters and non-teacher raters. Rating consistency and accuracy can be improved. Gundersen's (1996) small scale study (n=10) showed that significant improvement in rating could be had after raters received training via simple video-taped modules and carefully reading the training instructions. Therefore, in my study we had trainings and it become evident that rating accuracy improved with training. The rating consistency and accuracy is detailed in Chapter VI.

Conclusion

The preceding literature review included the areas of story retell, vocabulary, oral proficiency, curriculum-based assessment, and psychometric considerations of languages tests and rating. The literature review manifests a general consensus in the research findings in each area. However, the same body of research made it evident that the research lacked coherence in terms of second language learning. In 2000, Read stated that the amount of research on second language acquisition had increased; but, that the field still lacked coherence. In 2006, Genesee et al (p. 226) reaffirmed that there still exists a need for coherent research:

> Widespread application of research findings to the benefit of large numbers of ELLs is more likely to come from sustained research efforts whose primary aim is a full and in-depth understanding of an issue rather than from one or two isolated studies on a specific topic. Applied research consisting of single-studies is not as useful as theory-driven research identifying the needs of ELLs across the United States.

CHAPTER III

METHODOLOGY

The first purpose of my study was to develop and validate the scores of a curriculum-based assessment measure for expressive vocabulary and oral proficiency: Semantic and Syntactic Scoring System (S4) (see Appendix A). The S4 was used to analyze the responses provided on the Project STELLA Vocabulary Fluency Measure (see Appendix B) by the kindergarten students in the large-scale project English Language and Literacy Acquisition (ELLA) (Lara-Alecio et al., 2003) . As part of the development and validation process teacher utility was also considered. The second purpose of my study was to use the S4 instrument and other commercial measures such as the language and vocabulary subsections of the Woodcock Language Proficiency battery-Revised (WLPB-R) and language and vocabulary subsections of the Iowa Test of Basic Skills (ITBS) to compare the performance of students who partook in instruction under the two most common bilingual education models: Transitional Bilingual Education (TBE) and Structured English Immersion (SEI), with control and experimental treatments for each under the Project ELLA.

The research procedures for Project ELLA and also my study, which used data from Project ELLA, were approved by the Institutional Review Board at Texas A&M University. In this chapter, I will first address the development and the validation of the S4. I continue with the following traditional sections which pertain to the S4 and the second part of the study collectively, as such:

sampling, research design, intervention, instrumentation, data collection, and data analysis, which are relevant to both the first and second parts of this study.

Research Design

To answer questions one and two, the research design used was correlational. To answer question three the research design for this study was a quasi-experimental, 2x2 factorial design. All data used were archived data from Project ELLA (Lara-Alecio, Irby, & Mathes, 2006) . In Project ELLA, 24 elementary schools that had existing TBE and SEI programs in place were part of the initial random selection. ELLA's design included 12 schools, each of which received an enhanced treatment. In these 12 schools, enhanced treatment, 10 schools received enhanced SEI and TBE, and the two remaining schools received either enhanced SEI or TBE. There were 12 control schools used in the project. In the 12 control schools, nine schools received unaltered SEI and TBE, and the remaining three schools had unaltered SEI.

Setting and Participants

The data used in this study were archived data from the first year of implementation of Project English Language and Literacy Acquisition (ELLA) (R305P030032)[1], a federally funded grant by the U.S. Department of Education (Lara-Alecio et al., 2003). The grant was a collaborative research project among three universities and one school district. The universities were Texas A&M University (TAMU), Sam Houston State University (SHSU), and Southern

[1] Data were archived data from existing data sets provided under the U.S. Department of Education, Institution of Education Science federal grant, Project ELLA, R305P030032.

Methodist University (SMU). Project ELLA's main purpose was to develop, implement, and evaluate programs that would improve English proficiency and reading achievement for kindergarten through third-grade students and to evaluate the efficacy and impact of those programs under the two most ubiquitous ELL education models: structured English immersion and transitional bilingual education.

Setting

The school district where the ELLA project was implemented was located in Texas, and herein, this district will be denominated with an alias: School District T. In 2006, School District T was recognized with the Texas Award for Performance Excellence, (TAPE). This award is given to Texas organizations that demonstrate excellence in performance and outstanding quality. Since the inception of this award in 1994, only one other school district had obtained this achievement (Texas Education Agency, 2006b). This large urban district is one of the three largest in Houston and was selected because it was deemed a reputable district as evidenced by being recognized 7 of the last 8 years prior to the beginning of my study and the TAPE award. According to the Texas Education Agency (2004), over half of the student population in District T were Hispanic, thus, it services large numbers of Spanish-speaking ELLs as reflected in Table 1.

Table 1

Ethnic Distribution of Students in District T and in Texas

Student Groups	District T	Texas
Hispanic	32,565 **58%**	1,868,318 43.8%
African-American	18,573 33.1%	614,714 14.3%
White	3,614 6.4%	1,669,842 38.7%
Native American	50 0.1%	13,752 0.3%
Asian/Pacific Islander	1,325 2.4%	126,875 2.9%

Note. Retrieved from the Texas Academic Excellence Indicator System website report 2003-2004, Texas Education Agency (2004)

District T partnered with project ELLA to evaluate the academic progress of 905 Hispanic kindergarten ELLs enrolled in language development programs. The 905 students were divided into two groups: the control and experimental. During the 2004-2005 academic school year the control group consisted of 20 ESL and 12 TBE classrooms that delivered instruction under the typical guidelines and regulations of the district. The experimental group consisted of 14 ESL and 12 TBE classrooms that incorporated the instructional model interventions defined by the grant. These interventions were classified into two

categories: Tier 1 Teacher Enhancement and Tier II Student Intervention. Both experimental and control elementary campuses were randomly selected to participate in Project ELLA. The reason that there were more ESL classes participating in Project ELLA was due to the small number of Hispanic ELLs in those classrooms. In District T, the ESL classroom typically consisted of students who spoke an array of different languages. Moreover, not all students in an ESL classroom are labeled as Limited English Proficiency (LEP). On the other hand, the majority, if not all, of the students enrolled in a TBE classroom were Hispanic and considered LEP.

The data were collected from 48 kindergarten classrooms (24 TBE and 24 SEI) among 12 elementary schools. In order to participate, the school had to have both SEI and TBE programs in place. The ELLA researchers selected 12 schools that had, at least, 2 SEI and 2 TBE classrooms at the kindergarten level.

Participants

In District T, 45% of the ELL students were serviced through structured English immersion, transitional bilingual, or two-way immersion and they spoke Spanish as their L1. These students were identified as limited English proficient as per state criteria after their parents/guardians indicated that the primary language spoken at home was Spanish. According to Texas Education Agency (2004) 81.3% of the students in the district were classified as low SES; thus, they were on free or reduced lunch. State law (Texas Education Agency, 1995)

mandates that Spanish-speaking students identified as having low English proficiency be placed in bilingual classrooms. However, parents are able to opt out of bilingual services by signing a waiver. Those parents who opted out of traditional bilingual classrooms had their children placed in SEI classes.

The Texas Education Code (1995) disallows random program placement of students. Therefore, the ELLA grant principal investigators applied a robust matching scheme to create language and literacy-equivalent groups. In order to mitigate the confines of program placements, the ELLA researchers identified match scores on the IDEA oral language proficiency test for each set of students. Children who did not have an equivalent match were eliminated from the sample. During the 2004-2005 academic school year, the students were placed in one of two programs: a late-exit transitional program (TBE) or an English as a Second Language program (ESL). Furthermore, the demographics of the students in each program type were similar in terms of school, socio-economic status, and culture.

Table 2 shows the number of kindergarten Hispanic ELLs that participated in Project ELLA during the 2004-2005 academic school year. It is evident that the number of students in the TBE experimental surpasses the numbers in other groups. The reason that there were more students in the experimental TBE was that two experimental teachers provided instruction to three different groups of students enrolled in a TBE classroom – adding a total of four additional groups to the experimental TBE groups.

Table 2

Kindergarten ELL Participants in Project ELLA, 2004-2005

	N for Experimental (Enhanced) 11 Schools	*N* for Control (Typical) 12 Schools	Total *N*	Total Classrooms
SEI	198	203	401	12 SEI enhanced 16 SEI typical
TBE	303	201	504	17 TBE enhanced 11 TBE typical
Total per Group	501	404	905	

Procedures

Under Project ELLA, a series of assessments were administered to the kindergarten-level participants of this study to establish baselines and measure progress in oral language proficiency and literacy. The assessments that were relevant to this study were scores from the STELLA Vocabulary Fluency Measure (which was scored using the S4), WLPB-R (language and vocabulary subtests), and ITBS (language and vocabulary subtests). These data were collected during the 2004 – 2005 academic school year.

Paraprofessionals and district substitutes facilitated the administration of many of the measures used in Project ELLA. The individuals involved with testing in project ELLA, underwent a three-day training in preparation for administering assessments to the students throughout the year. The examiners were trained on the importance of adhering to the testing procedures as indicated in the testing manuals in order to ensure fidelity of test scores. The examiners were given opportunities to practice giving the assessments that they would administer and would not be allowed to proceed with testing until they demonstrated competency, which was part of the check-out process with a program coordinator. This check-out process required the test administrator to simulate the test and to accurately deliver the instrument and recording of scores.

Program Intervention

The following reiterates the differences in language instruction and intervention among the program types in Project ELLA (Lara-Alecio et al., 2003). There were four program types in Project ELLA: Control Structured English Immersion (SEI-C), Control Transitional Bilingual Education (TBE-E), Experimental Structured English Immersion (SEI-E), and Experimental Transitional Bilingual Education (TBE-E).

Control Structured English Immersion

Control SEI was the typical SEI program currently taught in District T. Under the current model, all subjects were taught in English with few

clarifications made in Spanish. The curriculum was aligned with the state reading and English as second language standards, the Texas Essential Knowledge and Skills.

Control Transitional Bilingual Education

Control TBE in District T, was the typical TBE program which they begin in kindergarten with an 80/20 language model: 80% in Spanish and 20% English. By grade 3, the model progresses to 50/50. The focus during kindergarten is oral language development in English, and progresses by grade three to content area instruction in English in subjects such as Science and Social Studies. The curriculum is aligned with the Texas Essential Knowledge and Skills.

Experimental Structured English Immersion

This program model was developed as part of the Project ELLA study. The SEI model is an enhanced version of the typical SEI used in District T. Under this model, English instruction was given with only minor clarifications in Spanish. The curriculum was also aligned with the TEKS. Under this model, teachers participated in weekly staff development opportunities in various areas such as: (a) enhancing instruction via planning, (b) support for student involvement, (c) vocabulary building and fluency, (d) oral language development, (e) literacy development, including the use of technology, (f) reading comprehension, and (g) parental support and involvement. The

paraprofessionals who worked under this model are also trained to work with students with an Intensive English program.

Experimental Transitional Bilingual Education

This program model was developed as part of the Project ELLA study. The TBE model is an enhanced version of the typical TBE model used in District T. Under this model, in kindergarten, language use is 70/30: 70% Spanish and 30% English. This distribution changes to a 40/60 model by grade three. This curriculum is aligned with the TEKS. In kindergarten the focus is English oral language development, which develops into Science and Social Studies for English content area instruction by grade three. Teachers were taught to use content as a tool to improve oracy, literacy, vocabulary and comprehension. Under this model, teachers also participate in weekly staff development opportunities in various areas: (a) enhancing instruction via planning, (b) support for student involvement, (c) vocabulary building and fluency, (d) oral language development, (e) literacy development, including the use of technology, (f) reading comprehension, and (g) parental support and involvement. Para professionals are trained to provide intensive daily English instruction for the students.

STELLA Intervention

*S*tory-retelling and higher order *T*hinking for *E*nglish *L*iteracy and *L*anguage *A*cquisition (STELLA) (Irby et al., 2008) was created as a critical intervention for the TBE and SEI experimental groups in project ELLA. This

intervention was systematically developed to serve as a structured and interactive story reading, pedagogical literacy intervention. The program used prior knowledge to enhance learning, literacy, and L2. Some of the scientifically-based pedagogical strategies that STELLA incorporated were: scaffolding, direct and indirect vocabulary instruction, higher-order thinking skills, interactive instruction, and question generation. The entire program was scripted and was executed in 5-day cycles (see Appendix M). One of the key elements of the program was the use of L1 clarifications (Lara-Alecio & Parker, 1994) that facilitate learning for second language learners. STELLA enhanced instruction provides a platform for the development of English academic or decontextualized language. The curriculum designers of STELLA realized as Pappas and Pettegrew (1991) found that teachers are using story retelling because they contribute to a holistic literary experience; they represent oral compositions and occasions for children to reconstruct or reenact text read, and they assist teachers in assessing students' comprehension.

STELLA Oral Proficiency

STELLA provided opportunities for students to respond with elaborate speech via oral language stems and questions that were a component of the instructional design. In a recent study STELLA was found to be an effective educational component in terms of improving young ELLs oral language development (Irby et al., 2008; Quiros, 2008; Tong, Lara-Alecio et al., 2008b).

STELLA Vocabulary

One of the ELLA researchers and a grant coordinator, who worked specifically with the STELLA intervention, systematically selected 18 vocabulary words from various books used in the STELLA intervention. The 18 words were *school, face, hop, climb, mittens, caterpillar, born, feathers, woods, scarf, munch, swooped, spring, crowd, squirm, shelter, perch,* and *trail*. Another ELLA Principal Investigator reviewed the words for accuracy and made final approval. Selection of words relied heavily on judgment, and in this case the bank from which the words were randomly selected was comprised of vocabulary words based on the Baumann and Kame'enui (2004) word selection suggestions. These suggestions are as follows: (a) words found in the instructional material, (b) words that can be defined at a kindergarten level, (c) words that are both useful and interesting, (d) words that are important for comprehension, and (e) words that might be known by the student but the student might not have a full understanding of, or the various ways in which the word can be used. Furthermore, the instrument facilitates the measuring of (a) the total time that it takes for the student to complete the assessment, (b) how many correctly used words are provided by the student in one minute, (c) the total words in correct sentences, and (d) the time factor for the production of the total number of correct words used in sentences. However, measurement of those factors was not within the scope of this study.

The vocabulary development, as well as the other aspects of STELLA, adhered to research on best practices. For example, the Reading Panel Report (National Institute of Child Health and Human Development, 2000a) conducted a comprehensive review of studies in vocabulary instruction. The report delineated the following recommendations for optimal vocabulary acquisition (STELLA reflected the recommendations): (a) vocabulary should be taught both directly and indirectly; (b) repetition and multiple exposures to enhance vocabulary learning; (c) it is important to learn in rich contexts; incidental learning is also encouraged; (d) students should be actively engaged, and (e) multiple vocabulary instructional method should be used. Vocabulary instruction is a key element of the STELLA Curriculum. Vocabulary is taught through direct and indirect instruction in conjunction with critical thinking skills. The target words that are taught in the STELLA intervention are defined, word usage is modeled, and the vocabulary word is practiced in and out of context. In kindergarten three words are introduced per day (Irby et al., 2008; B. J. Irby et al., 2004).

STELLA Comprehension

In terms of enhancing comprehension, STELLA reflects best practice as deemed by the National Reading Panel (2000), which identified seven comprehension strategies that have scientific basis for improving comprehension and they are as follows: comprehension monitoring, cooperative learning, use of graphic and semantic organizers (including story maps),

question answering, question generation, story structure use, and summarization (see Appendix M). The panel found that it was most effective for these strategies to be combined, as opposed to being taught in isolation.

Instrumentation

In Project ELLA there were sundry testing instruments used to measure student progress. In this section, I first address the development of the Semantic and Syntactic Scoring System (S4), the impetus of my study. The S4 was an instrument developed to analyze the oral sentences produced by kindergarten ELLs in the administration of the STELLA Vocabulary Fluency Measure Protocol (add citation) which was developed under Project ELLA. Then I provide details on two extant, norm-referenced, and commercial measures: The Woodcock Language Proficiency Battery – Revised (WLPB-R) (Woodcock, 1991a) and the Iowa Test of Basic Skills (ITBS) (Hoover, Hieronymus, Frisbie, & Dunbar, 2006).

Development of the S4

This section describes the methodology and the systematic process of instrument development for the S4. The first purpose of this study was to create the Semantic and Syntactic Scoring System (S4) for the Project STELLA Vocabulary Fluency Measure Protocol (add citation here). The Project STELLA Vocabulary Fluency Measure Protocol is a curriculum-based, criterion-referenced measure that attempts to effectively measure vocabulary knowledge through students' ability to use words in oral sentences. A researcher-created archetype from project OPTIMIZE (PacifiCorp Foundation, 2004) was used by

the STELLA intervention designers in Project ELLA, to inform in the development of the protocol. The S4 was developed through my study to analyze student responses in the Project STELLA Vocabulary Fluency Measure.

My study, in instrument development, resembles the study by Howard, Christian, and Genesee (2004) in which they used a researcher-developed proficiency measure to evaluate English proficiency and Spanish proficiency. However, in that study no information was provided on the reliability or validity of the scores for the measure. The examiners in that study were representatives from each of the schools that participated in the study. These 12 representatives/examiners received a two-day training with the oral proficiency assessment that the researchers modeled after a writing rubric. After the training, a researcher visited each school and administered the instrument along with the trained school representative/ examiner. Only a subsample took this oral proficiency test. In third-grade, the subsample size was 247 students. In fifth-grade, the subsample size was 234 students. Students were interviewed in pairs (paired according to similar proficiency levels, as determined by the teachers). The test administration lasted 15 minutes. The researcher provided the students with social and academic prompts and students were allowed to help each other and ask questions of test administrators. The school representative acted as the examiner and the researcher rated the student's performance as the student spoke. The researcher recorded the students' responses by writing them down and tape recording the session (to review for

questionable scores, revisiting, & for more substantive future analyses). The categories measured were conversational fluency, comprehension, fluency, vocabulary, and rhetorical complexity. In terms of grammar, verbs, verb agreement, word placement, and prepositions were measured. The scale was a 6 point scale (0-5). An average of 8 subcomponents was used to obtain a total score. Only total scores were discussed in the report. Native English speakers were compared to non-native English speaking Two-Way instructed students. In the third-grade there was some variability of scores, the *f-statistic* was 56.27. However, the instrument did not do an effective job at distinguishing students in the fifth-grade, there was very little variability of scores, and the *f-statistic* was 12.13.

Project STELLA Vocabulary Fluency Measure

The intent in creating the S4 for the Project STELLA Vocabulary Fluency Measure Protocol, a researcher-created archetype modified from the DIBELS measure Word Use Fluency – Grades K and First (University of Oregon Center on Teaching and Learning, n.d.), was to create a criterion-referenced ruler that effectively measures vocabulary knowledge through students' ability to use words in oral sentences.

Characteristics of Effective Language Instruments

Effective language instruments should adhere to the following: (a) construct validity in terms of oracy and vocabulary development, (b) yield sufficient variability (differentiation) among the levels in the measure, as well as

judgments regarding quality and accuracy, (c) inclusion of generally accepted attributes of language in terms of expressive vocabulary and oral proficiency, (d) yield generalizable results through desirable psychometric properties of interrater reliability, (e) ability to detect subtleties among individuals and groups in order to afford proper assessment and ranking, and (e) permit efficient observation, scoring, and use of the measure (North, 2000; Read, 2000).

S4 Scale Descriptors

The descriptors for the S4 were developed a priori and based on the theoretical underpinnings of vocabulary and oral proficiency as delineated in the Literature Review of this study. Also, two studies' (i.e.,Eller, Pappas, & Brown, 1988; Leung & Pikulski, 1990) scales were used to provide guidance in developing the scale descriptors for the S4. Leung and Pikulski used the following descriptors (see Table 3) to rate vocabulary use in a pretest and posttest.

Table 3

Leung and Pikulski's Descriptors for Vocabulary Analysis

Descriptors for a vocabulary test which were used to anchor the descriptors for the S4.	
0 points	No knowledge of word meaning or incorrect response
1 point	Partial or incomplete knowledge of word meaning
2 points	Target word used in an appropriate, meaningful context
3 points	Synonym or definition of target word

Table 4

Eller, Pappas, and Brown's Descriptors for Vocabulary Analysis

Eller created these descriptors for a vocabulary test based on a reading intervention and these were also used to anchor the descriptors for the S4.

Category One (No/Faulty Knowledge)	Indicates no knowledge or a faulty knowledge of the word's meaning. a) Target word was not used. b) A non-synonymous replacement was used
Category Two (Developing Knowledge)	Indicates developing knowledge of semantic and syntactic features of the word, but knowledge still seems incomplete or faulty. a) Target word was used, but used inappropriately of contained a syntactic error. b) Target word was used inappropriately elsewhere in the text.
Category Three (Synonym)	Indicates that the child has obtained semantic and syntactic information about the word from context, but is still using a more familiar word to impart his/her message. a) Synonyms word or phrase used. b) When the word occurred more than once in the text, child supplied a synonym as frequently as the target word.
Category Four (Accurate Knowledge)	Indicates not only an acquisition of accurate semantic and syntactic information about the word, but also that this information may be internalized so that the target word is now used appropriately within the given context. a) Accurate use of target word in given context. b) Accurate use of target word elsewhere in the text, but not in conjunction with (a). c) When the word occurred more than once in the text, child supplied the target word more frequently than a synonym.
Category Five (Generalized Knowledge)	Indicates that generalization may have occurred in that the word was used accurately in both given and other contexts within the text. a) Accurate use of the target word not only in given context, but also elsewhere in the text, use of target word in given context.

Eller et al. (1988) created a system to analyze target word knowledge during a reading intervention and it is summarized in Table 4.

Iterations of the S4

The S4 underwent five iterations. In the initial iteration researchers, Irby, Pollard-Durodola, and I decided what discrete elements of language could be measured in terms of vocabulary and oral proficiency, decided on the levels of the scale, created a four levelscale, and defined the descriptors for the scale. Then after the initial development, the S4 underwent an additional four iterations in which the scale was tested and refined based on use and feedback with individuals not directly involved with Project ELLA, STELLA, or the S4.

Initial S4 Development. In the initial development of the S4 Scoring system, I worked with two researchers from the ELLA grant, Irby and Pollard-Durodola, in the process of creating descriptors that showed sufficient differentiation among the four levels. During this phase the scale did not have a 0 descriptor level; the scale range was from 1-4. I selected grade-level appropriate target words and produced sample sentences (similar to the types of responses that kinder ELLs at various levels of proficiency would be expected to produce on a measure such as the STELLA Vocabulary Oral Proficiency Measure). We rated these responses independently and then checked for percent agreement. Whenever we disagreed on the scoring, we would discuss the differences to determine the rationale behind each persons rating. Then, I modified the descriptors, so as to better differentiate among each level based on

theoretical and practical premises. As I improved the descriptions and differentiation of the descriptors, our percent agreement was consistently in the high 90s. This first iteration of the Semantic and Syntactic Scoring System (S4) is included as Appendix C. I proceeded to create the scale and materials to test for reliability with other researchers, graduate students, and in-service teachers.

Initial Scale Development of the S4

First Iteration of S4. The first iteration used the S4 which was produced in the first iteration and included training materials. The training packet consisted of a Semantic and Syntactic Scoring System (Appendix B1), Practice A: Distinguishing between levels 2 and 3 (Appendix C), Practice A: Distinguishing between levels 3 and 4 (Appendix D), Independent Practice with 30 items (Appendix E), and a Final Practice (Appendix F). These materials were used to train two ELLA researchers and three graduate students to use the S4 accurately. The training duration was 1.5 hours. The raters were presented with the S4 and oral clarifications and discussion were provided on each of the descriptors. Then the raters attempted the Practices with checking and follow-up discussion on each. Interrater reliability and percent agreement were scored using the Final Practice. During the training, questions that the raters asked in terms of ambiguity in rating due to the descriptors and the rationale that they used to rate sentences were used to further modify the descriptor levels and produce another version of the S4.

Second Iteration of the S4. The second iteration of the S4 included a more detailed version of the 5 rating levels; it included 0 for No Response (see Appendix G) and a Progression Chart (Appendix H). The training materials used with the first iteration were also used in this second iteration. Furthermore, this iteration included three practices with actual data (see Appendix I, Appendix J, and Appendix K).This training was conducted with four coordinators involved in the ELLA grant. Three of the four grant coordinators were doctoral graduate students. All the ELLA coordinators had over five-years of elementary teaching experience in ELL settings. The training and rating session lasted 2 hours. Again, questions and comments posed by the raters in terms of the descriptors were considered and then used to inform the third iteration of the materials.

It became evident, in the second training session and iteration development, that the raters were influenced by language mazes and tended to rate students lower who were manifesting a language maze in their sentence production. Loban (1976, p. 74) defined a language maze as follows:

> In as much as fluency connotes flow of language, its success can be marred by too many hesitations, false starts, and nonfunctional repetitions. Because the language tangles very much resemble the physical behavior of a person seeking a way out of a maze, we called them mazes at the beginning of our research, and the name stuck. We defined maze as a series of words (or initial parts of words), or unattached fragments which do not constitute a communication unit and are not necessarily to the communication unit. It is only in speech that these language tangles occur, and if one listens attentively to anyone's oral language, or indeed one's own, it soon becomes apparent that the phenomenon is universal. Obviously, it appears to be related to the problems of putting thought and feeling into words, what might be called verbal planning. In writing, one can pause as long as desired, crossing out extraneous words of bugled phrases, thus eliminating mazes.

Raters were informed that they were to rate without the influence of the mazes that children produced and to determine if a sentence had been produced within the maze. An example of a maze would be, "*Uhm, ahh*, the *uh*, boy *like*, likes to munch on the *uhm* carrots." Some raters were distracted by these maze-like phrases presented in italics and would tend towards penalizing the student for the mazes by giving them a score that was one or two points lower than the sentence warranted. After this became evident subsequent training and feedback sessions I clarified that mazes were not to influence scoring. Information about the language mazes was also added to the training and scoring materials.

Third Iteration of the S4. In the third iteration all the training and scoring materials were used. The raters for this iteration were three doctoral students in educational psychology. These doctoral students served as new raters; these raters did not participate in any previous or subsequent iterations. The training and rating in this session was just under two hours. For these raters it seemed that most confusion lay between the descriptors 2 and 3 and 3 and 4. There was a concern over rater behaviors that could affect reliability such as what Meltzoff (1998, p. 98) called "…rater drift, fatigue, boredom, flagging of attention, and loss of interest and motivation." In an effort to improve the confusion among the descriptors and to decrease the effects of negative rater behavior, I created a

Progression Chart (Appendix L), which was to be used with actual data and allowed for focused and better directed rating of the sentences produced.

Fourth Iteration of the S4. The fourth and final iteration of the S4 included a detailed scoring manual (Appendix M). The intent in this final iteration was to create a manual and materials that teachers could use without needing to be trained in-person. These materials were sent to four elementary, bilingual teachers. They reviewed the manual and did a practice rating. They were able to compare their scores on the Practice sheet with explanations of the correct answers. Then the teachers scored an Independent Practice which was used to evaluate percent agreement and interrater reliability. As with all the former rating procedures, the raters could use and were encouraged to use their Scoring Summaries, charts, and any other information provided to assist them in scoring the sentences.

Woodcock Language Proficiency Battery – Revised

The Woodcock Language Proficiency Battery – Revised (WLPB-R) (Woodcock, 1991b) is a set of individually-administered assessments for non-native speakers of English used to measure ability and proficiency in oral language, reading, and written language. The revised version contains modifications that increase the diagnostic applicability of the instrument. Traditionally, this instrument is used for diagnosis, program placement, establishing Individual Education Plans, educational guidance, assessing growth over time, program assessment, and research. Normative data for the WLPB-R

was gathered from 6,359 subjects in over 100 U.S. communities. Internal consistency estimates for the subtests and clusters were all in the .80s and .90s. Test-retest reliability was in the .70s and .80s. Concurrent validity was analyzed with other instruments such as the Boehm Basic Concepts, bracken Basic Concepts, Stanford-Binet IV, and the WISC – R and in general the correlation coefficients with other measure ranged from the .30s to .70s (National Clearing House for English Language Acquisition and Language Instruction Educational Programs, n.d.). The instrument is attributed with providing an inclusive measure of English language competence and that is why it was selected as a measurement instrument in the ELLA project and as a measure for correlation with the S4. For this study the scores on the following subtest were the only ones deemed relevant: picture vocabulary, verbal analogies, and listening comprehension.

Relevant Subtests. In the Picture Vocabulary test children are asked to identify pictures of familiar and unfamiliar objects. As the test progresses the objects that are depicted become less familiar. As part of this test there is a word retrieval component. In Verbal Analogies, students complete a logical word association. The words in this section of the test are simple, but the relationship between the words increases in complexity as one progresses through the test. Listening Comprehension is another subtest and it requires students to listen to a story and provide a single-word response to a cloze-type statement.

ITBS

The purpose of the Iowa Test of Basic Skills (ITBS) (Hoover et al., 2006) is to provide an indicator of progress for students in major content areas in order to facilitate instructional, such as curricular decisions and placement decisions in grades K-8 (levels 5-14). This instrument was normed in 2000 and 2005. It is a group-administered test and takes 30 minutes or less per test. Separate scores are provided for each section in diagnostic reports of strengths and weaknesses. Vocabulary, Word Analysis, Reading Comprehension, Language, Math, Social Studies, Science, in general are the sections offered across the levels. Herein, I have reported relevant information for the sections that the kindergarten participants of this study took and those that are relevant in measuring the construct of language proficiency in terms of oral language and vocabulary.

Relevant Subtests. Vocabularies presented in the test are general vocabulary words. This section measures the 'overall breadth' of students' vocabulary and is an indicator of overall verbal ability. Receptive vocabulary is the focus of levels 5 and 6. Students identify the appropriate one of three pictures upon hearing a word used in a sentence. Word Analysis is focused on phonological awareness and morphology. In levels 5 and 6 students identify letter and sound relationships. Level 5 is pictorial and level 6 begins to introduce some word responses. Listening comprehension is tested in levels 5 through 9. These are scenarios that are orally presented with subsequent questions. The

test measures the following comprehension abilities: understanding, following directions, visualizing objects, making inferences, understanding concepts and sequence, and predicting outcomes. Language tests at levels 5 through 6 measures the ability of students in expressing ideas. The skills that are assessed are the use of prepositions, comparative and superlatives, and singular plural distinctions. Here again, scenarios are presented in a picture format and students choose the picture that indicates a correct response.

Research Questions

The following four questions guided this study:

1. To what extent can a curriculum-based assessment instrument be developed and validated to measure oral proficiency and vocabulary knowledge among ELLs who are participating in a story retell intervention?
2. Can teachers use the Semantic and Syntactic Scoring System (S4) for the STELLA Vocabulary Fluency Measure with minimal training to accurately assess students' vocabulary knowledge and oral proficiency?
3. To what extent does the developed curriculum-based assessment instrument, Semantic and Syntactic Scoring System (S4) differentiate the level of knowledge regarding expressive vocabulary and oral proficiency of kindergarten students participating in the STELLA intervention under two different programs: enhanced Traditional Bilingual Education and the enhanced Structured English Immersion Program in comparison to the Revised Woodcock Language Proficiency battery (WLPB-R) (language and vocabulary subtests)?

Data Collection

The data for this study were archival data collected during the regular course of the ELLA grant in District T. These data were retrieved after the Institutional Review Board of Texas A&M University granted permission to execute this study and with permission from the PIs in the grant.

Assessment Schedule

The data that were used for this study were scores from the S4, WLPB-R, and ITBS for the participants of this study from the academic years, 2004-2005. The timeframe for these exams are depicted in Table 5 (see Assessment Schedule for Project ELLA, 2004-2005).

Table 5

Assessment Schedule for Project ELLA, 2004-2005

	Beginning Fall	Mid Spring	End Spring
S4		√	
WLPB-R	√		√
ITBS		√	

Commercial Measures

The data from the WLPB-R and the ITBS were collected during the regular course of the ELLA grant in 2004-2005. Trained paraprofessionals or testers administered these tests. The data were collected using a Tele-form software. The data were entered and cleaned under Project ELLA. As per IRB conditions the data for this study were provided using fictitious identification numbers for the students. However, the identification number used for each student was consistent across all measures.

Curriculum-based Assessment

The Project STELLA Oral Proficiency measure was administered to each student, individually, by a trained paraprofessional. The testing of 813 participants was completed in six weeks, during the spring of 2005. The students were taken out of the classroom for this test. Children met with a paraprofessional in a quiet room in their respective school. Each student was instructed in English and Spanish that they were to provide a sentence using the words they were given. The administrator provided two examples. In the first example, the examiner provided an example, such as “If I say *run*, you might say, ‘the dog *runs* along the beach.’” “Now it is your turn: *cat*.” The student then used the word *cat* in a sentence. If the sentence was grammatically and semantically correct, the examiner affirmed by saying “good job.” If the student did not provide a response, merely repeated the word, or provided an erroneous response, the examiners modeled the correct response by saying, “You could

have said, ‘We give milk to the cat.’” Then the students were provided with a similar second example. After the second example, students were provided with target words from the STELLA curriculum. All the students were given the same 18 words: *school, face, hop, climb, mittens, caterpillar, born, feathers, woods, scarf, munch, swooped, spring, crowd, squirm, shelter, perch, and trail*. The examiner would give the word and pause for 30 seconds in order to allow the student to think and provide a response. If the student responded by repeating the word or by providing a sentence, then the examiner proceeded to the next word. If the student did not give a response, then the examiner asked five more words. If the student did not respond to those five consecutive words then the examiner would stop the test. At the end of the test, students were thanked for their participation and sent back to class. Each student administration took between one minute and five minutes, for most students. If the students merely repeated the word, the administration took less than a minute (usually 40 - 50 seconds). If the student provided simple sentences, the administration of the probe took between 2-3 minutes. It was only when the students provided extended elaboration or needed much time to think and construct a sentence that took between 4 and 5 minutes to complete the probe.

Trained paraprofessionals administered this measure and recorded the entire protocol and examination onto tapes. The examiners, carefully, labeled each tape for each given class with the teacher’s name. The teacher of record was also mentioned at the beginning of the test on the tape. The examiners

provided the students' name at the beginning of each tape and class. Only one tape was used per class, whether there were five students in the group or over 20. The students were identified at the beginning to ensure that the student responses were credited to the correct student. The examiners used a class roster which included the teacher's name, student name, and identification number. Most of the time the examiners tested students in alphabetical order, as their last names appeared on the roster. However, if a student was absent or not available then they deviated from the order. The students that were absent were tested on a make-up day which was within a couple of weeks from their scheduled testing.

These rosters were kept in a clear Poly-See through string envelope. Each classroom roster was stored in an individual envelope under secure conditions in the Project ELLA office. In the envelope there were copies of the STELLA Vocabulary Fluency Protocol for each student. These were used during the transcribing phase. The tapes, upon completion, were stored in the respective envelope to facilitate the transcription process which took place during the Summer of 2005.

Three individuals provided the transcriptions of five samples, and comparisons were made for interrater reliability. It seems that there were, at times, technical problems with the audio recordings. Ten children did not speak directly into the recorders, and thus there were discrepancies with regards to what the transcribers heard. Forty-three students were tested in an environment

that had much background noise, and that also hindered interrater reliability. Therefore, 53 recordings were not used in the data. Also, at times there were inconsistencies in being able to decipher word endings, which is the area the raters differed more than in any other area. Difficulties in understanding children's speech (diction and pronunciation) are endemic with second language learners.

Data Analysis

The results of the WLPB-R (language and vocabulary subtests), ITBS (language and vocabulary subtests), and STELLA Vocabulary Fluency Measure using the (S4) were gathered, coded, entered for analysis into SPSS versions 15. Each student was assigned an identification number, which was consistent for the student across all measures.

Descriptive statistics were completed for the raw scores S4, WLPB-R (language and vocabulary subtests), and ITBS (language and vocabulary subtests). These data were presented based on program type: Transitional Bilingual Education control and experimental and Structured English Immersion, control and experimental.

Then the scores on S4, WLPB-R subtests, and ITBS subtests were tested for normality in terms of visual analysis, skewness, and kurtosis. In particular the normality assumptions were evaluated closely for the S4 because that was the instrument of interest. Parametric and mixed model analyses were employed where appropriate. The data in this study were naturally nested:

within schools, classrooms, and programs, as is the case with much educational research. Therefore, multilevel models, such as hierarchical linear models were used for some of the analyses. Mixed Models (Allison, 1999; Bryk & Raudenbush, 1992; Demidenko, 2004) allow the researcher to take into account the school, campus, and teacher effects. Using a Mixed Model analysis in this study facilitated going beyond fixed effects into random effects, so this research can be generalized to different campuses and different teachers. And a Mixed Model will handle the mixed effects of the fixed effect and the random effects. Furthermore, it is helps with missing data which did occur with 96 of the 909 possible participants. Some of the missing data were due to inaudible recordings, and the rest were due to absences on the day of the S4 administered or attrition. The alpha level was set at .05 for all analysis because the nature of this study was exploratory. Additionally, effect sizes were calculated.

Summary

Chapter III included the Methodology for this study and the following: Development of the S4, Research Design, Sampling, Program Intervention, STELLA Intervention, Instrumentation, Research Questions, Data Collection, setting, research design, instrumentation, intervention procedure, data collection, Data Analysis, and Summary. Chapter IV includes the Results.

CHAPTER IV

RESULTS

In this chapter, I included data exploration results and the findings for the three research questions. The research questions for this study were:

1. To what extent can a curriculum-based assessment instrument be developed and validated to measure oral proficiency and vocabulary knowledge among ELLs who are participating in a story retell intervention?
2. Can teachers use the Semantic and Syntactic Scoring System (S4) for the STELLA Vocabulary Fluency Measure with minimal training to accurately assess students' vocabulary knowledge and oral proficiency?
3. To what extent does the developed curriculum-based assessment instrument, Semantic and Syntactic Scoring System (S4) differentiate the level of knowledge regarding expressive vocabulary and oral proficiency of kindergarten students participating in the STELLA intervention under two different programs: enhanced Traditional Bilingual Education and the enhanced Structured English Immersion Program in comparison to the Revised Woodcock Language Proficiency battery (WLPB-R) (language and vocabulary subtests)?

The raw scores of the Semantic and Syntactic Scoring System (S4), Woodcock Language Proficiency Battery (WLPB-R) subtests (Picture Vocabulary, Listening Comprehension, and Verbal Analogies), and Iowa Test of

Basic Skills (ITBS) subtests (vocabulary and word analysis), were gathered, coded, entered for analysis into SPSS® version 15. Each participant was assigned an identification number which was consistent for that student across all measures.

First, descriptive statistics were completed. The mean, standard deviations, and ranges were calculated for all the measures: S4, WLPB-R subtests, and ITBS subtests.

Then the scores on S4, WLPB-R subtests, and ITBS subtests were tested for normality in terms of visual analysis, skewness, and kurtosis. Parametric analysis was used to answer questions one and two, and mixed model (hierarchical linear model) analyses were employed to answer question three. Furthermore, the data in this study were naturally nested: within schools, classrooms, and programs. Multilevel models, such as hierarchical linear models, are designed for data which are naturally clustered into groups (Bryk & Raudenbush, 1992; Kreft & De Leeuw, 1998) thus, mixed model (aka Hierarchical linear model) analysis was also used.

The alpha for achieving statistical significance was set at .05 for all analyses. Although an alpha of .05 results in an experiment error rate that is greater than stated alpha level, this study was an exploratory study, and it was important to use the same alpha across multiple tests. Thompson (2006, p. 304) defined experimentwise error rate, (α Experimentwise) as referring to:

> …the probability of having made one or more Type I errors anywhere within the study. When only one hypothesis is tested for a given group of

> participants in a study, the experimentwise error rate will exactly equal the testwise error rate. But when more than one hypothesis is tested in a given study, the two error rates may not be equal.

Additionally, effect sizes were included. The results of the analyses are presented by research question after the Data Exploration section.

Data Exploration

The data for my study included scores on three measures. The first measure was the Woodcock Language Proficiency Battery – Revised (WLPB-R) which has three subtests that pertain to this study because they are focused on language and vocabulary. The three relevant WLPB-R subtests are Picture Vocabulary (PV), Listening Comprehension (LC), and Verbal Analysis (VA). This standardized and validated instrument was administered in the fall 2004 and the spring 2005; thus, providing pretest and posttest data. The second, instrument was the IOWA Test of Basic Skills (ITBS). The ITBS has two relevant subtests: Vocabulary (VO) and Word Analysis (WA). The ITBS was only administered in the spring 2005; therefore, the scores for this instrument are treated as posttests. The third instrument is the impetus of this study and is the Semantic and Syntactic Scoring System (S4). The S4 was only administered in the spring 2005; therefore, the scores for this instrument are treated as posttests.

Table 6 depicts the pretest scores on the three WLPB-R subtests (PV, LC, and VA) for descriptive purposes. These scores will be used to create a covariate score for the S4 because the S4 did not have a pretest score.

Table 6

Descriptive Statistics for the Pretest Scores on the WLPB-R Subtests

	n	Range	Mean	Std. Dev.	Skewness	Std. Error	Kurtosis	Std. Error
PV	819	0 - 33	14.19	5.552	-.151	.085	-.012	.171
LC	808	0 - 20	3.56	4.064	1.211	.086	.778	.172
VA	797	0 - 11	1.57	1.877	1.369	.087	1.975	.173

Note. These Pretest data were collected in the Fall 2004. PV=WLPB-R (Picture Vocabulary), LC=WLPB-R (Listening Comprehension), VA= WLPB-R (Word Analysis), VO= ITBS (Vocabulary), WA= ITBS (Word Analysis).

Table 7 depicts the posttests scores for the S4 and for the subtests for the WLPB-R and the ITBS. The S4 was of upmost consideration since it is the impetus of this study. The S4 had values of skewness (.212, SE.086) and kurtosis (-0.547. SE .171). The S4 tended towards a bimodal distribution and that influenced the skewness and kurtosis statistics.

Table 7

Descriptive Statistics for Posttests and Subtest Measures

	n	Range	Mean	Std. Dev.	Skewness	Std. Error	Kurtosis	Std. Error
PV	816	0 - 31	19.51	4.236	-.439	.086	1.254	.171
LC	817	0 - 25	6.00	4.989	.625	.086	-.371	.171
VA	816	0 - 18	2.60	2.286	1.369	.086	3.622	.171
ITBS VO	807	0 - 27	14.50	3.890	-.169	.086	1.390	.172
ITBS WA	797	1 - 4	3.95	.345	-7.045	.087	50.441	.173
S4	813	0 - 69	23.14	15.075	.212	.086	-0.547	.171

Note. These Posttest data were collected in the Spring 2005. PV=WLPB-R (Picture Vocabulary), LC=WLPB-R (Listening Comprehension), VA= WLPB-R (Word Analysis), VO= ITBS (Vocabulary), WA= ITBS (Word Analysis), and S4=Syntactic and Semantic Scoring System.

Most of the statistics for the posttests on Table 7 are have small variation among the different measures, except for ITBS Word Analysis. ITBS WA has a the largest kurtosis (50.441). The range of scores for the ITBS WA was between 1 and 4. And 85.6% of the scores accounted for a score of 4.

Furthermore, with large sample sizes of 200 or more, “it is more important to look at the shape of the distribution visually and to look at the value of skewness and kurtosis rather than calculate [the standard error] significance” (Field, 2005, p.72). Figure 3 provides a visual illustration of the distribution of scores on the S4 for the four groups: Transitional Bilingual Education control (TBE-C), Transitional Bilingual Education experimental (TBE-E), Structured English Immersion control (SEI-C) and Structured English Immersion experimental (SEI-E).

The distributions for the S4 in Figure 3, appear to be bimodal. This is inherent of and consistent with the testing protocol for the STELLA Vocabulary Fluency Measure and the S4 scoring rubric. If students did not give a response when they tested, their score was a 0. The test was administered in English and required students to create an English sentence for the target words. Another data point that seems to occur frequently among all groups is the score of 18. This occurred because students scored a 1 for each item if they repeated the word and did not provide a sentence. Many students would just repeat each of the 18 words and not give a sentence which would give them a raw score of 18.

Once the test administration was in progress, the examiner could not redirect the student to create a sentence.

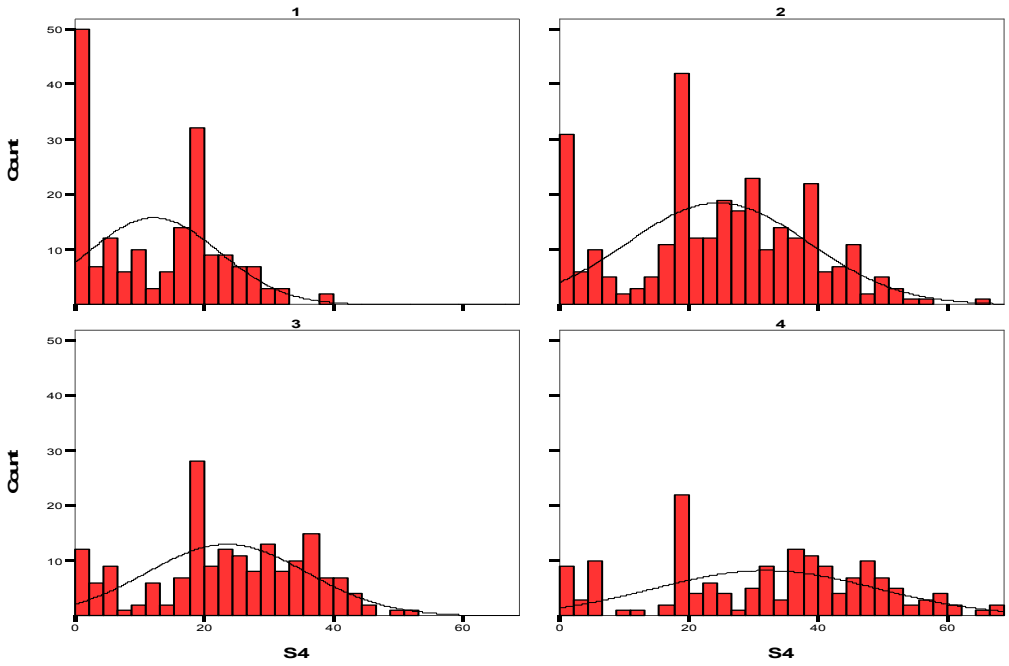

Figure 3. Frequency graph of the S4 with normality curves. 1=TBE control, 2= TBE experimental, 3=SEI control, and 4=SEI experimental.

Since the WLPB-R was the only measure that was administered as a pretest and posttest, the score for each subtest WLPB-R subtests is documented on Table 8 by group (TBE-C, TBE-E, SEI-C, SEI-E), for descriptive purposes. In each instance, the experimental enhanced treatment groups made a greater improvement than the control groups. The pretests scores of the SEI groups were higher than those of the TBE group across the subtests.

Table 8

Means of WLPB – R Pretest and Posttests by Group

	TBE Control WLPB-R n = 173		TBE Experimental WLPB-R n = 291		SEI Control WLPB-R n = 175		SEI Experimental WLPB-R n = 173	
	Pre	*Post*	*Pre*	*Post*	*Pre*	*Post*	*Pre*	*Post*
PV	11.63 (sd 5.28)	16.55 (sd 4.03)	12.33 (sd 4.72)	17.89 (sd 3.26)	18.03 (sd 5.18)	22.67 (sd 3.09)	16.17 (sd 4.58)	22.28 (sd 3.08)
Δ	4.92		5.56		4.64		6.11	
LC	1.72 (sd 2.42)	3.36 (sd 3.66)	2.02 (sd 2.82)	4.22 (sd 3.88)	6.34 (sd 4.67)	9.16 (sd 5.03)	5.26 (sd 4.30)	8.65 (sd 4.88)
Δ	1.64		2.2		2.82		3.39	
VA	.91 (sd 1.40)	1.81 (sd 1.86)	1.24 (sd 1.66)	2.18 (sd 1.77)	2.51 (sd 2.30)	3.44 (sd 2.74)	1.86 (sd 1.74)	3.29 (sd 2.51)
Δ	0.9		0.94		0.93		1.43	

Note. PV=WLPB-R (Picture Vocabulary), LC=WLPB-R (Listening Comprehension), VA= WLPB-R (Word Analysis), Δ = Change score (Posttest-Pretest).

Since the level of scale for the WLPB-R subtests is different for each subtest a better comparisons of pretest and posttest performance by group can be determined through Cohen's *d* and the effect size *r.* Cohen's *d* is the difference between two means, divided by the standard deviation of either group (Cohen, 1988). Table 9 depicts the Cohen's d and effect size statistics which

indicated that the there were initial group differences and that the experimental groups outperformed the control groups under the Typical Transition Bilingual Education (TBE) and Structured English Immersion (SEI) models. It was apparent that the enhanced SEI group had the greatest gain from pretest to posttest.

Table 9

Cohen's d and Effect Size for the WLPB – R Pretests and Posttests by Group

	TBE Control WLPB-R n = 173	TBE Experimental WLPB-R n = 291	SEI Control WLPB-R n = 175	SEI Experimental WLPB-R n = 173
	d	*d*	*d*	*d*
PV	1.05	1.37	1.09	1.57
LC	0.529	0.649	0.581	0.737
VA	0.547	0.548	0.368	0.622

Note. PV=WLPB-R (Picture Vocabulary), LC=WLPB-R (Listening Comprehension), VA= WLPB-R (Word Analysis), d= Cohen's d and r = effect size.

Results for Research Question 1

Research question 1: to what extent can a curriculum-based measure be developed and validated to measure oral proficiency and vocabulary knowledge among ELLs who are participating in a story retell intervention? To answer this question Pearson's *r* was used to assess the concurrent validity of the S4 with the picture vocabulary, listening comprehension, and verbal analogies

subsections of the WLPB-R posttests and vocabulary and word analysis subsections of the ITBS posttests at the kindergarten level. Also, Cohen's Kappa and percent agreements were calculated to ascertain the reliability of the S4 in terms of intrarater reliability and interrater reliability.

Validity

The testing standards developed by the American Psychological Association (APA) and the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME) convey that validity is the most fundamental consideration in developing and evaluating tests and the process of validating a measure involves the compilation of evidence that provides a scientific basis for the interpretation of the scores on a given measure (1999, p. 9). The most important issue in language testing is that of validity because a test needs to measure what it purports to measure (Alderson, Clapham, & Wall, 1995a).

Therefore, this first research question addressed concurrent validity which can be tested by evaluating statistical evidence to see whether students scores on a given measure are similar to the scores obtained on other appropriate and comparable measures. These can represent scores on tests, self-assessment, or even teacher ratings of ability (Alderson, Clapham, & Wall, 1995b). For this analysis I used the Pearson product-moment correlation coefficient (*r*) (Sheskin, 2007) a widely used correlation indicator in the social sciences (Bachman, 2004). A correlation coefficient is a measure of the

relationship between two variables. And it is also an index of the proportion of individual differences in one variable associated with the proportional differences of another variable (Hinkle, Wiersma, & Jurs, 2003). Only the posttests of the WLPB-R and ITBS were used in this analysis, because they were concurrently administered with the S4 . There is little point in comparing students' test scores with their performance on some measure that is not reliable or valid (Alderson et al., 1995a); therefore, the reliability and validity of the WLPB-R and ITBS were detailed in Chapter III and proved to be adequate and these are tests that are regularly used in school districts.

Table 10 depicts the interpretation criteria for interpreting the magnitude of correlation coefficients from little correlation to very high (positive or negative) correlation as provided by Hinkle et al (2003). This standard was used to evaluate the data in this study.

Table 10

Table for Interpreting Correlation Coefficients

Size of Correlation	Interpretation
.90 to 1.00 (-.90 to -1.00)	Very high positive (negative) correlation
.70 to .90 (-.70 to -.90)	High positive (negative) correlation
.50 to .70 (-.50 to -.70)	Moderate positive (negative) correlation
.30 to .50 (30 to .50)	Low positive (negative) correlation
.00 to .30 (00 to -.30)	Little if any correlation

Note: (Hinkle, Wiersma, Jurs, 2003, p. 109)

Table 11 illustrates the correlation results for the S4 with the WLPB-R subtests (Picture Vocabulary, Listening Comprehension, and Vocabulary Analysis) and ITBS subtests (Vocabulary and Word Analysis). The range of the correlation coefficients for the S4 compared to the WLPB-R and ITBS relevant subtests are on the low to moderate end (.133 to .457); however, all the coefficients when compared to the S4 were statistically significant ($p<.01$) and directionally all positive, as expected. The coefficients that were not statistically significant were the correlations of the two subtests of the ITBS (Vocabulary and Word Analysis). The range of the coefficient of determination (r^2) is between .021 to .210, which means that the proportion of the variance in Y (S4) that can be associated with the variance in X (the language and vocabulary subtests) is less than .30 which as a rule of thumb indicates that there is minimal relationship (see Table 10) among the S4 and the extant, standardized, and validated measures that are used in ELLA: WLPB-R and ITBS.

Table 11

Correlation Coefficients for the S4, WLPB-R, and ITBS Posttest Measures

	S4	PV	LC	VA	VO	WA
PV	**.445****					
LC	**.457****	.652**				
VA	**.374****	.480**	.599**			
VO	**.283****	.437**	.394**	.340**		
WA	**.133****	.209**	.134**	.078**	**.058**	

Note: ** p<0.01 (2-tailed), * p<0.05 (2-tailed). These Posttests data were collected in the Spring 2005. PV=WLPB-R (Picture Vocabulary). LC=WLPB-R (Listening Comprehension). VA= WLPB-R (Word Analysis) VO= ITBS (Vocabulary), WA= ITBS (Word Analysis), and S4=Syntactic and Semantic Scoring System.

Table 12 depicts the correlation analysis results for the 4 groups (TBE-C, TBE-E, SEI-C, and SEI-EI), for the S4 with WLPB-R (Picture Vocabulary, Listening Comprehension, and Vocabulary Analysis), and ITBS (Vocabulary and Word Analysis).

In Table 12 the range of the coefficients is on the low end: .023 to .523; however, most of the coefficients are statistically significant (p<.01), and they are directionally positive. As in Table 11, the measures that did not always have statistical significance were the correlations of the subtests of the ITBS with the S4. ITBS Vocabulary correlations with the S4 were not statistically significant in

the TBE-C group. The ITBS Word Analysis correlations with S4 were not statistically significant in either SEI-C group nor the SEI-E group.

Table 12

Correlation Coefficients for the S4 with WLPB-R and ITBS by Group

	TBE Control WLPB-R n = 173 ITBS n =159	TBE Experimental WLPB-R n = 291 ITBS n =290	SEI Control WLPB-R n = 175 ITBS n =172	SEI Experimental WLPB-R n = 173 ITBS n =159
PV	.336**	.509**	.378**	.241**
LC	.439**	.451**	.357**	.413**
VA	.385**	.380**	.292**	.326**
VO	.157	.161**	.228**	.310**
WA	.197*	.143**	.025	-.003

Note: ** p<0.01, (2-tailed), * p<0.05, (2-tailed). These Posttests data were collected in the Spring 2005. PV=WLPB-R (Picture Vocabulary). LC=WLPB-R (Listening Comprehension). VA= WLPB-R (Word Analysis) VO= ITBS (Vocabulary), WA= ITBS (Word Analysis), and S4=Syntactic and Semantic Scoring System.

Correlations are maximized when each of the variables being correlated is normally distributed, with good dispersion of scores. This allows maximum opportunity for variation in one measure to be associated with variation on the other. If one or both of the variables is restricted in range then the correlations obtained will usually be lower (Skehan, 1989).

The correlation coefficients for these data were low, yet statistically significant, perhaps due to the large *n* in this study (Field, 2005; Hinkle et al., 2003; Skehan, 1989). Notwithstanding this concern, studies using large sample sizes enhance the reliability of their findings (Hinkle et al., 2003). Furthermore, it is important to evaluate the interpretation of correlation coefficients in respect to the context of the data because Skehan (1989, p .13) conveyed that, "in practice, second language learning studies yield correlations whose maximum values rarely approach +1 and are more likely to be in the order of 0.30 – 0.60."

Reliability

The testing standards developed by the American Psychological Association (APA) and the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME) (1999, p. 25) defined reliability as "…the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups." Fulcher and Davidson (2007) imparted that in classical test theory there are three assumptions in the concept of reliability and these assumptions are as follows: (a) the person or environment being tested has a true score, which is some fixed

amount of the attribute being observed, (b) all observations of an attribute contain some degree of error and (c) the observed score contains both true and error variance. In answering the first research question reliability is considered in terms of intrarater reliability and interrater reliability.

Intrarater Reliability. Intrarater reliability was calculated at seven different time points for the researcher. The intrarater check was a rescoring of a student probe that had already been scored by the researcher, previously. Each student probe provided 18 sentences (one sentence produced by the student for each of the 18 target words) and these were the items used to calculate reliability between the first time the student's test was scored and the second time it was scored, after a period of time (2 weeks, 1 month, or 4 months apart). The seven student probes were randomly selected using the random selection feature in SPSS Version 15 and coded for the second scoring. These randomly selected probes were copied twice. The first copy of the probe was used in the first scoring. The second copies were placed in a separate folder for the second rescoring, thus, ensuring blindness to the previous scoring. The first two intrarater reliability checks occurred while developing the scale and were within two weeks apart. The third and fourth intrarater checks were conducted after the scale was changed to include five descriptors (5-point-scale) instead of four descriptors (4-point-scale) and these interrater reliability rescoring were conducted one month apart. The last three (fifth, sixth, and seventh) intrarater

checks were conducted after the scale was finalized into a flow chart form (see Appendix L) and these were conducted approximately four months apart.

> An examiner is judged to have *intra-rater reliability* if he or she gives the same set of scripts or oral performances the same marks on two different occasions. The examiner may still be considered reliable even if the marks are different; However, not much variation can be allowed before the reliability becomes questionable. Intra-rater reliability is usually measured by means of a correlation coefficient or through some form of analysis of variance. (Alderson et al., 1995b, p. 129)

The seven intra-rater reliability checks, as depicted in Table 13, reflect that the intra-rater reliability coefficients and percent agreement improved as the S4 scale descriptors were refined at each iteration. Although, Cramer' V and Kappa are traditionally used to calculate chance-corrected agreement between two raters, in this case they were employed to calculate the chance-corrected agreement of the researcher at two different occasions.

Table 13

Intrarater Reliability Correlation Coefficients, Effect Sizes, and Percent Agreement

	Cramer's V	*Kappa*	%
First (2 weeks apart)	.339	.817	83%
Second (2 weeks apart)	.330	.837	83%
Third (1 month apart)	.565	.911	83%
Fourth (1 month apart)	.825	.923	89%
Fifth (4 months apart)	.914	.913	100%
Sixth (4 months apart)	1.00	1.00	100%
Seventh (4 months apart)	1.00	1.00	100%

Note: Each statistic corresponds to reliability based on rating 18 items from a given students randomly selected student probe, (18 scored sentences) by the researcher after some passage of time. With each iteration a different student probe was randomly selected for scoring and archived for the second rescoring.

Cramer's V is interpreted as a measure of relative strength of an association between two variables. It is not affected by sample size and can be treated as an adequate effect size since it is constrained to fall between 0 and 1 which makes it easy to interpret (Acock & Gordon, 1979; Field, 2005). Cohen's Kappa is a chance-corrected measure of association which is used to quantify the agreement between two judges. In general the value of measures of association or correlation tend to range between 0 and +1 or -1 and +1; whereas, 0 indicates no relationship and 1 indicates a perfect relationship (Cohen, 1960; Sheskin, 2007). Table 14 (Altman, 1991) presents a possible

interpretation of agreement. In intrarater reliability there was an improvement from fair agreement to consistently good and very good agreement.

Table 14

Table for Interpreting Kappa

Size of Correlation	Interpretation
Less than 0.20	Poor Agreement
0.20 to 0.40	Fair Agreement
0.40 to 0.60	Moderate Agreement
0.60 to 0.80	Good Agreement
0.80 to 1.00	Very Good Agreement

Note: (Altman, 1991, p. 404)

Interrater Reliability. Not only does intrarater reliability address the issue of reliability for the S4, it was important to establish high interater reliability because the researcher's scoring was used as the chief examiner or standard for comparison with the other raters in establishing the descriptors and in conducting interrater reliability checks after the instrument was finalized. It was evident that the researcher was able to achieve *good* and *very good* agreement which justifies consistency in the scale use and is indicative of being able to set the standard for interrater reliability checks. Interrater reliability was also

evaluated to determine if raters could distinguish between the descriptors on the S4.

> Though there is bound to be some variation between examiners and the standard some of the time, there must be a high degree of consistency overall of the test is to be considered reliable by its users…reliability is measured by a correlation coefficient or by some form of analysis of variance. (Alderson et al., 1995b, p. 129)

Just as with intrarater reliability, Cramer's V (Acock & Gordon, 1979; Field, 2005; Sheskin, 2007) and Kappa (Cohen, 1960; Field, 2005; Sheskin, 2007) were used to calculate interrater reliability, the agreement among raters and the researcher (which was the standard). Also, percent agreement was calculated overall for all the raters and individually. Overall (general) Kappa was calculated as opposed to using the arithmetic mean for all possible paired-rater Kappas. King (2004) stated that using the arithmetic mean for all possible paired-rater Kappas is the equivalent of averaging multiple t-tests rather than conducting an analysis of variance and that perhaps the failure to use a generalized kappa stems from the omission of generalized kappa in most statistical computing packages Furthermore, King (2004) developed a Microsoft Excel spreadsheet, which was used for the multiple rater analyses in this study, it was based on the estimates of the generalized kappa statistics as proposed by Fleiss (1971; 1981) and discussed by Berry and Mielke (1988).

During the first iteration of the S4, the raters that were used to test the distinctiveness and clarity of the descriptors were as follows: raters 1, 2, and 3 were doctoral students in educational psychology and raters 4 and 5 were

professors in education (and they were directly involved with the ELLA grant and the STELLA intervention). I, as the researcher of this study, was Rater 6. In the first iteration, each rater was compared to the standard (the researcher) to obtain the statistics for Table 15, Overall (general) Kappa was calculated using all 6 raters in comparison to each other. Overall Kappa was .587. Overall percent correct was calculated among the first five raters in comparisons to Rater 6 (the standard). Overall percent correct was .40. For the first iteration the range of the Cramer's V statistic was .594 to .951 (not including the researcher) and they are statistically significant ($p<.01$). Also, the range of Kappa was .431 to .953 (not including the researcher). According to Table 14, these Kappa coefficients are considered *moderate* to *very good* agreement.

Table 15

Interrater Reliability Statistics for First Iteration

	Kappa	Cramer's V	% Correct
	Overall Kappa: .587		Overall: 40
Rater 1	.475	.616	37%
Rater 2	.431	.594	43%
Rater 3	.645	.739	70%
Rater 4	.953	.951	97%
Rater 5	.953	.951	97%
Rater 6	1.00	1.00	100%

Note: *p<0.01, (1-tailed), *Rater 6 was the researcher*

The raters in the second iteration were grant coordinators in Project ELLA. Raters 7, 8, and 9, and 10 were doctoral students in educational psychology or curriculum and instruction. Rater 8 contributed to the STELLA curriculum. The researcher was Rater 11. In the second iteration, each rater was compared to the standard (the researcher) to obtain the statistics for Table 16. Overall (general) Kappa was calculated using all five raters in comparison to each other. Overall Kappa was .728. Overall percent correct was calculated among the first five raters in comparisons to Rater 11 (the researcher/standard). Overall percent correct was .60. The percent correct of each rater in comparison to the standard ranged from .70 to .87 (not including the researcher). For the second iteration the range of the Cramer's V statistic was .736 to .856 (not including the researcher) and they are statistically significant ($p<.01$). Also, the range of Kappa was .682 to .860 (not including the researcher). As demonstrated in Table 14, these Kappa coefficients are considered *moderate* to *very good* agreement.

Table 16

Interrater Reliability Statistics for Second Iteration

	Kappa	Cramer's V	% Correct
	General Kappa: .728		Overall: 60
Rater 7	.817	.856	83%
Rater 8	.860	.900	87%
Rater 9	.682	.739	70%
Rater 10	.855	.869	83%
Rater 11	1.00	1.00	100%

*Note: *p*<0.01, (1-tailed), *Rater 5 was the researcher*

The raters in the third iteration were all doctoral students in educational psychology and were different from those that participated in previous iterations. The researcher was Rater 15. In the third iteration, each rater was compared to the standard (the researcher) to obtain the statistics for Table 17. Overall (general) Kappa was calculated using all four raters in comparison to each other. Overall Kappa was .809. Overall percent correct was calculated among the first five raters in comparisons to Rater 15 (the researcher/standard). Overall percent correct was .72. The percent correct of each rater in comparison to the standard ranged from .72 to .77 (not including the researcher). For the third

iteration the range of the Cramer's V statistic was .700 to .769 (not including the researcher) and they are statistically significant (p<.01). Also, the range of Kappa was .682 to .860 (not including the researcher). According to Table 14, these Kappa coefficients are considered *moderate* to *very good* agreement.

Table 17

Interrater Reliability Statistics for Third Iteration

	Kappa	Cramer's V	% Correct
	General Kappa: .809		Overall: 72
Rater 12	.817	.700	72%
Rater 13	.860	.769	77%
Rater 14	.682	.769	77%
Rater 15	1.00	1.00	100%

*Note: *p*<0.01, (1-tailed), *Rater 5 was the researcher*

With each iteration the statistics for overall general Kappa, overall percent correct, Cramer's V, and between raters Kappa, improved. The fourth iteration was not used to change the descriptors of the S4. The fourth iteration tested the use of the flow chart format and the training was done using only the manual and not in-person training. The fourth iteration answers question 2 and is detailed in that section. Even in comparison to the fourth iteration, these

interrater reliability figures were to closer to 1 which indicate a greater consistency in rating (Hock, 2003).

Summary

The section for Research Question 1, showed that in terms of concurrent validity the correlation of the S4 with WLPB-R language and vocabulary subtests and ITBS language and vocabulary subtests were low. However, the S4 is a curriculum-based measure and is purported to provide a specific measure of ability that would not be provided in extant, standardized, and commercial measures. Second, this section included reliability of the S4 by measuring intrarater and interrater reliability. Intrarater reliability (which was the researchers consistency in rating) and interrater reliability (which was the consistency of scoring for other raters in comparison to each other and the standard which was set by the researcher) improved as a function of three things: (a) the wording in the descriptors were changed to be more specific, (b) the scale was modified from a 4 point scale to a 5 point scale, (c) improvements in training, and finally (d) the S4 format was changed to facilitate scoring in a sequential and decision-making flow-chart style.

Results for Research Question 2

Research question 2: can teachers use the Semantic and Syntactic Scoring System (S4) for the STELLA Vocabulary Fluency Measure with minimal training to accurately assess student performance? This question addressed two aspects of instrument development: reliability and utility. To answer this

question overall general Kappa, Overall percent correct, Cramer's V, and between raters Kappa analyses were used. Finally, descriptors are presented to address utility of the S4.

The raters in the fourth iteration were all bilingual-elementary-school teachers. The researcher was Rater 19. The teachers did not receive any in-person training. They were emailed the training manual, the S4 flow chart, and the probes used for this interrater reliability check. The teachers scored the interrater reliability probes and sent them back to the researcher via email. All the teachers worked on this independently. The teachers were from 2 different schools. In this fourth iteration, each rater was compared to the standard (the researcher) to obtain the statistics for Table 18. Overall (general) Kappa was calculated using all four raters in comparison to each other. Overall Kappa was .812. Overall percent correct was calculated among the first three raters in comparisons to Rater 4 (the researcher/standard). Overall percent correct was .72. The percent correct of each rater in comparison to the standard ranged from .83 to 1.00 (not including the researcher). The range of the Cramer's V statistic was .822 to 1.00 (not including the researcher) and they were statistically significant ($p<.01$). Also, the range of Kappa was .786 to 1.00 (not including the researcher). According to Table 14, these Kappa coefficients are considered *good* to *very good* agreement.

Table 18

Interrater Reliability Statistics for Teacher Raters in Fourth Iteration

	Kappa	Cramer's V	% Correct
	General Kappa: .812		Overall: 72
Rater 16	1.00	1.00	100%
Rater 17	.792	.860	83%
Rater 18	.786	.822	83%
Rater 19	1.00	1.00	100%

Note: *p<0.01, (1-tailed), *Rater 4 was the researcher*

In answering question two it was important to take into account the time commitment on behalf of the teachers for reading the S4 training manual and to score an 18-target word probe. Table 19 depicts the time that it took each bilingual elementary school teacher to read the manual and to score one 18-target-word probe (the researcher is not included in the table). The time to read the manual ranged from 10 to 30 minutes (μ=17.5, SD 8.66). The scoring of a single probe (Scoring 1 and Scoring 2) took between 9 and 35 minutes (μ=19.38, SD 8.81).

Table 19

Time in Minutes Expended by Teachers Using the S4 Manual and Self-training Materials

Raters	Reading Manual	Scoring 1	Scoring 2	Total Time
Rater 16	10	15	15	40
Rater 17	30	35	30	95
Rater 18	15	9	21	45
Mean Time	18.3	19.6	22	

It might take a teacher 9 to 35 minutes to rate a student's response for an 18 target word probe. Which if an average class size is 20 students this could mean that a teacher would spend between 3 to 12 hours using this instrument to score all the student in his or her class, not including the training. For the researcher reading the manual was not necessary. As the researcher having the page open that provided the descriptors was the only thing that needed to be reviewed which took less than one minute. Scoring time for the researcher per student was 5 minutes per probe or less. For the researcher to rate a classroom of 20 students it would take just over 1.5 hours and this is consistent with the time spent rating the 814 assessments (which were divided, in most cases, into classes of 18 students).

Summary

The section for Research Question 2, included reliability and utility results for the S4 with three bilingual, elementary teachers, raters. These teachers raters did not receive any in- person training. They used the S4 manual and scoring chart to rate a student probe. Since the teachers received the final iteration of the S4 scale and flowchart, it is important to note that their interrater reliability (in terms of Cramer's V, Kappa, and percent correct) were overall higher even though they were not trained in-person to use this instrument. This section also provided information in terms of time used to score a student probe. The range was between 9 to 35 minutes. However, this was a one time rating and did not take into consideration a learning effect, which would make scoring faster and more efficient with time and practice.

Results for Research Question 3

Research question 3: to what extent does the developed curriculum-based assessment instrument, Semantic and Syntactic Scoring System (S4) differentiate the level of knowledge regarding expressive vocabulary and oral proficiency of kindergarten students participating in the STELLA intervention under two different programs: enhanced Transitional Bilingual Education and the enhanced Structured English Immersion Program in comparison to the Revised Woodcock Language Proficiency Battery (WLPB-R) (language and vocabulary subsets)? To answer this question a Mixed Model Regression was used because it is a robust analysis which can take into account the nested nature of

my data. Pedhazur (1982, p. 34) affirmed that, “It has been demonstrated that regression analysis is generally robust in the presence of departures from assumptions, except for measurement errors and specification errors.”

The design of this study was dictated by the initial design of the Project English Language Literacy Acquisition (ELLA) project (R305P030032), a federally funded grant by the U. S. Department of Education (Lara-Alecio et al., 2003) . This longitudinal project used a quasi-experimental design because of the nature of schools and ELL program placement, which do not permit that students be randomly assigned to schools and programs. The data for this study were the kindergarten archived data for the first year of the project. There were 905 kindergarten participants who were divided between control and experimental in a TBE or SEI program. The data were collected from 48 kindergarten classrooms (with a different teacher for each of these 48 classrooms) among 12 elementary schools. Because the participants in Project ELLA, whose kindergarten data were used in this study, were not randomized to campus or teacher, the campus and teacher effects were considered as random effects in order to better generalize beyond the participants and settings of this particular study. And in this case, as with other research in educational settings a simple regression model should not be used “because the variables for students in a given classroom are considered correlated because for a variety of reasons, students in the same classroom tend to be more alike in academic performance than students in different classrooms. The consequences of

violating this assumption are standard errors that are too low and tests statistics that are too high" (Allison, 1999, p.182). Therefore, to answer this question, which utilizes nested data, a Mixed Model Regression was used. Mixed Model methodology has many names (model for repeated measures and hierarchical model) and many applications (i.e. analysis of clustered, panel, or longitudinal data) (Demidenko, 2004, p. 1).

> Mixed model methodology brings statistics to the next level. In classical statistics a typical assumption is that observations are drawn from the same general population, are independent and identically distributed. Mixed model data have a more complex, multilevel, hierarchical structure. Observations between levels or clusters are independent, but observations within each cluster are dependent because they belong to the same subpopulation. Consequently, we speak of two sources of variation: between clusters and within clusters.

Again, as established in Chapter III, the Mixed Model tested campus and teacher and were entered to allow for the effects of the nesting. Then the individual effects were accounted for by using the pretest scores from the WLPB-R. After accounting for the effect of campus, teacher, and student's beginning language ability (as measured by the WLPB-R) the next step was to see what differences there were in the S4 scores for the participants of the 4 groups: TBE-C, TBE-E, SEI-C, SEI-E.

WLPB-R Subtests as Covariate of S4

Since the students in this study did not have pretest scores for the S4, it was important to account for students' beginning levels of English oral proficiency and vocabulary knowledge by using pretests scores from other measures that reflected the same construct that S4 measures and use those

scores as covariates (substitutes for the S4 pretest). In Project ELLA, the WLPB-R was used to test all the participants and this test is deemed to provide an inclusive measure of English language competence, as discussed in Chapter III. Furthermore, the WLPB-R was administered in the fall and spring, thus providing a pretest and posttest score for the participants. Specifically 3 subtests form the WLPB-R were used because they related to the construct that the S4 is attempting to measure. The WLPB-R subtests were Picture Vocabulary, Listening Comprehension, and Vocabulary Analysis and these were used as predictors of S4. Table 20 depicts the overall pretest scores on the WLPB-R subtests for descriptive purposes.

Table 20

Descriptive Statistics for Pretest on WLPB-R (Subtest Measures)

	N	Range	Mean	Std. Dev.	Skewness	Std. Error	Kurtosis	Std. Error
WLPB-R PV	819	0 - 33	14.19	5.552	-.151	.085	-.012	.171
WLPB-R LC	808	0 - 20	3.56	4.064	1.211	.086	.778	.172
WLPB-R VA	797	0 - 11	1.57	1.877	1.369	.087	1.975	.173

Note. These Pretest data were collected in the Fall 2004. PV=WLPB-R (Picture Vocabulary), LC=WLPB-R (Listening Comprehension), VA= WLPB-R (Word Analysis).

The scores on the WLPB-R pretest Picture Vocabulary, Listening Comprehension, and Vocabulary Analysis were analyzed to see if the assumption that they could be used as covariates for the S4 was correct. The regression analysis shows that the multiple r was .465. The r-squared was .216 which indicates that the three pretest scores account for 22% of the variance in S4. Therefore, the relationship between the WLPB-R pretests scores correlate with the post test S4 roughly to the same degree that they each correlate with their own posttests. For Picture Vocabulary the bivariate r was .598 with an r-squared of .357 or 35% variance. The bivariate r for the Listening Comprehension subtest was .675 and the r-squared was .456 or 46% variance. And the bivariate r for Vocabulary Analogies was .455 and the multiple r-squared was .207 or 21% variance.

Testing for Random and Fixed Effects of Campus and Teacher

The gain from a Mixed Model regression is that it can handle the nesting, but at the expense of needing to specify the correct correlation structure a priori. The analysis does not provide that. In order to do that the researcher specified what the researcher thought should be in the model and then ran it with a different correlation structure and the information criteria from both were compared. The model with the lower values in the Information Criteria was deemed to be the model of best fit. Also, it is important to note that the Mixed Model analysis includes calculated variance components, spherisity, and autoregression within the analysis. These statistics impact and adjust the

degrees of freedom in the analysis. That empirical decision-making process is what is presented in this section. First campus was tested to see if it was a fixed or random effect. And the estimate of the covariance parameter for campus was 8.750 with a standard error of the estimate of 3.930. And for teacher the covariance parameter was 7.56 and the standard error of the estimate was 4.69. And one way to determine statistically if they are different from zero is to look at the ratio of the estimate to its standard error. The values of those estimates are, which can be considered quasi-z-scores, were 2.23 for campus and 1.61 for teacher with standard errors above 0. Then the Information Criteria, such as 2 Restricted Log Likelihood (Wolfinger, Tobias, & Sall, 1994), Akaike Information Criteria (Akaike, 1974), Hurvich and Tsai's Criterion (AICC) (Hurvich & Tsai, 1989), Bozdogan's Criterion (CAIC) (Bozdogan, 1987), and Schwarz's Bayesian Criterion (BIC) (Schwarz, 1978) were evaluated to determine if the model should include just campus or campus in conjunction with teacher as random effects. The values of the information criteria, in each instance, decreased when the teacher effect was added as evident in Table 21.

Table 21

Information Criteria Used to Determine Model Fit with DV S4

Information Criteria	With Campus	With Campus and Teacher
2 Restricted Log Likelihood	5909.114	5893.959
AIC	5913.114	5899.959
AICC	5913.130	5899.991
CAIC	5924.357	5916.823
BIC	5922.357	5913.823

Furthermore, campus and teacher were considered random effects due to the nested nature of the design; however, it was helpful to test that assumption and determine if they should be treated as random effects or fixed effects in the analysis model. The variance estimates are presented in Table 22 and they are not zero. And one way to determine statistically if they are different from zero is to look at the ratio of the estimate to its standard error. The values of those estimates are, which can be considered quasi-z-scores, 1.61 for campus and 2.39 for teacher with standard errors above 0. This indicates that these factors needed to be treated as random instead of fixed effects.

Table 22

Estimates of Covariance Parameters for Campus and Teacher (Classroom) for S4

Parameters	Estimate	Std. Error
Residual	131.0169	7.019478
Intercept Variance Campus	7.564595	4.689913
Intercept Variance Teacher	12.307339	5.146775

Testing the Full Model with S4

Results of the mixed model regression (see Table 23) indicated a significant intervention (control versus experimental) effect. Analysis of interaction effects indicated that there was no significant interaction between program type (SEI and TBE) and intervention (control and experimental). The scores of the S4 for the students in the TBE-E (*M*= 24.67, *SD*=14.08) and SEI-E (*M*= 31.98 *SD*= 17.16) were higher than those of TBE-C (*M*= 12.33 *SD*= 10.12) and SEI-C (*M*= 23.65 *SD*= 12.37).

Table 23

Type III Fixed Effects with DV S4

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	238.088	42.505	.000*
Program Type (TBE/SEI)	1	588.496	3.651	.057
Intervention (Ctrl/Exp)	1	23.609	57.413	.000*
Interaction	1	529.279	.197	.657
Pre PV	1	686.645	27.038	.000*
Pre LC	1	750.624	31.514	.000*
Pre VA	1	750.280	10.482	.001*

Note: $p < 0.05$

Testing Full Model on WLPB-R Subtests

Therefore, by considering the information presented in the covariate section and the random and fixed effect section to establish that it was acceptable to use the 3 WLPB-R pretests (Picture Vocabulary, Listening Comprehension, and Verbal Analysis) as covariates and campus and teacher were determined to be random effects, then the analysis that proceeds, statistically equates for these factors when looking at program type (TBE or

SEI), intervention (control or experimental), and interaction of program type and intervention.

The next step was to test the model established (and used to analyze S4 and presented in Table 23), to see if the variables considered functioned the same for the WLPB-R subtests as they did for the S4. The following sections, present the full model applied to each of the WLPB-R subtests by examining the Type III fixed effects and the Estimates of Covariance Parameters (for campus and teacher) for each of the WLPB-R subtests: Post Picture Vocabulary, Post Listening Comprehension, and POST Vocabulary Analysis.

For Post Picture Vocabulary the results of the mixed model regression (see Table 24) indicated a significant program type (TBE versus SEI) effect. The intervention was not statistically significant. Analysis of interaction effects indicated that there was a statistically significant interaction between program type (SEI and TBE) and intervention (control and experimental). The scores of the WLPB-R Posttest Picture Vocabulary were as follows: TBE-E (*M*= 17.89, *SD*= 3.257), SEI-E (*M*= 22.28 *SD*= 3.079), TBE-C (*M*= 16.55 *SD*= 4.031), and SEI-C (*M*= 22.67 *SD*= 3.097).

Table 24

Type III Fixed Effects with DV WLPB-R (Post Picture Vocabulary)

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	238.879	1597.447	.000
Program Type (TBE/SEI)	1	67.052	80.916	.000
Intervention (Ctrl/Exp)	1	20.960	3.026	.097
Interaction	1	46.511	4.401	.041
Pre PV	1	629.925	98.804	.000
Pre LC	1	765.643	27.791	.000
Pre VA	1	765.979	20.947	.000

Results of the mixed model regression (see Table 25) for the WLPB-R posttest Listening Comprehension, indicated a statistically significant program type effect. Analysis of interaction effects indicated that there was no significant interaction between program type (SEI and TBE) and intervention (control and experimental) for the Listening Comprehension subtest. The scores of this subtest were as follows: TBE-E (*M*= 4.22, *SD*=3.885), SEI-E (*M*= 8.65 *SD*= 4.887), TBE-C (*M*= 3.36 *SD*= 3.633), and SEI-C (*M*= 9.16 *SD*= 5.032).

Table 25

Type III Fixed Effects with DV WLPB-R (Post Listening Comprehension)

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	231.997	14.313	.000
Program Type (TBE/SEI)	1	73.224	23.653	.000
Intervention (Ctrl/Exp)	1	23.915	.653	.427
Interaction	1	52.856	.036	.849
Pre PV	1	688.860	19.682	.000
Pre LC	1	765.242	146.190	.000
Pre VA	1	766.480	14.947	.000

For post Listening Comprehension, (as presented in Table 26), the estimate for campus was 1.49 and for teacher was 2.18, which was calculated by dividing the ratio of the estimate by the standard error for each. Listening Comprehension had the second highest variability in campus and teacher. The S4 had higher estimates under each.

Table 26

Estimates of Covariance Parameters for Campus and Teacher (Classroom) for WLPB-R Post Listening Comprehension

Parameters	Estimate	Std. Error
Residual	10.935057	.577185
Intercept Variance Campus	.517359	.346598
Intercept Variance Teacher	.769672	.353090

In Table 27 the results of the mixed model regression manifested a non-significant effects for program type effect. Analysis of interaction effects indicated that there was no statistically significant interaction between program type (SEI and TBE) and intervention (control and experimental). The scores of the Vocabulary Analysis posttest were as follows: TBE-E (*M*= 2.18, *SD*=1.774) and SEI-E (*M*= 3.29 *SD*= 2.514) were higher than those of TBE-C (*M*= 1.81 *SD*= 1.866) and SEI-C (*M*= 3.44 *SD*= 2.741).

Table 27

Type III Fixed Effects with DV WLPB-R (Posttests Vocabulary Analogies)

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	249.992	12.064	.001
Program Type (TBE/SEI)	1	54.091	.254	.616
Intervention (Ctrl/Exp)	1	19.890	1.432	.246
Interaction	1	37.687	.280	.600
Pre PV	1	586.135	6.700	.010
Pre LC	1	763.698	74.996	.000
Pre VA	1	765.412	24.744	.000

For post Vocabulary Analogies, (as presented in Table 28), the estimate for campus was 0.68 and for teacher was 1.61, which was calculated by dividing the ratio of the estimate by the standard error for each.

Table 28

Estimates of Covariance Parameters for Campus and Teacher (Classroom) for WLPB-R Post Vocabulary Analogies

Parameters	Estimate	Std. Error
Residual	3.338631	.177773
Intercept Variance Campus	.061570	.090028
Intercept Variance Teacher	.204067	.126449

Effect Size and Summary

Because this study was an exploratory study, there was no preconceived, preset effect size because it was important to see the difference in group performance. I did find adequate effect sizes, so it was evident that there was sufficient power.

The effect size used in this study was based on the pretests and posttests for the WLPB-R subtests (Picture Vocabulary, Listening Comprehension, and Vocabulary Analysis). The formula used for Effect sizes could not be calculated for the S4 because there were no pretest scores on S4 for the calculation. The effect sizes in educational research are standardized mean differences between the treatment and control groups to a standard deviation. In cluster-randomized trials there are several standardized possible

differences (Hedges, 2007). The formula chosen for effect size calculation was $\delta\Gamma \equiv \frac{\mu^{\Gamma} - \mu^{c}}{\sigma\Gamma}$ (Hedges, 2007). The effect size was 0.328118 which was 0.328 which can be considered fair based on Table 14 which interprets Kappa but can also be applied here.

In summary, Table 29 provides the effects of each measure as detailed in this chapter, but with Yes and No responses. The S4 was the only measure that distinguished between the differences attributed to the control and experimental groups. S4 was a curriculum-based measure and it was able to distinguish between the children that received the instruction, which was part of the curriculum measured by S4. This is an important attribute of the S4. The S4 also manifested an interaction effect for program type and control/experimental. The S4 could not distinguish between TBE and SEI programs. In referring back to Figure 4.1, the TBE placements (1 was control and 2 was experimental) both have similar score distributions, in terms of high frequencies with the score of 0 and 20. This means that the students tended to not answer anything or answer in Spanish and they received a score of 0 or they just repeated the word and that gave them a score of 18. It makes sense that with such young ELLs that this tendency would prevail. The SEI control and experimental groups (3 was control and 4 was experimental) had a wider distribution of scores, yet there were higher frequencies in the 0 and 18 range. Overall, there were fewer children that scored 0 or 18 with the SEI groups in comparison to the TBE

children. But these were still the most frequent scores among the SEI groups. This can be attributed to the age of the children, as well. It would be expected that children in the SEI groups would have more English proficiency than those in the TBE groups because the foci of the programs are different. TBE is focused on Spanish instruction during the majority of the day in kindergarten. SEI programs are focused on English instruction throughout the day.

Table 29

Summary of Effects by Measure

	S4	PV	LC	VA
Program Type (BIL/SEI)	No	Yes	Yes	No
Intervention (CTRL/EXP)	Yes	No	No	No
Interaction	Yes	Yes	No	No

In looking at means, as provided in Table 30, the difference between the control and experimental group is more than a 10-point difference. The means reflect the analysis provided the mixed model analysis. It is also evident that with the TBE there was a greater influence from the instruction. This would

make sense because the students in the TBE experimental program were instructed in English for longer periods of the day, than the control. The SEI experimental also manifested better scores than SEI control; however, since both of these groups were receiving instruction in English one would expect less of a difference between them when measured with an instrument that looks at English oral proficiency.

Table 30

Means on the S4 in Each Group

	Program Type		
	SEI	TBE	Total
Control Groups	23.65 (sd 12.37) (n=204)	12.33 (sd 10.12) (n= 204)	17.99 (n=408)
Experimental Groups	31.98 (sd 17.16) (n=184)	24.67 (sd 14.08) (n=317)	28.33 (n=501)
Total	27.815 (n=388)	18.5 (n=521)	

The next chapter, Chapter V, presents a discussion which incorporates the analysis presented in this chapter in light of the corpus of literature presented in Chapter II.

CHAPTER V

SUMMARY, DISCUSSION, IMPLICATIONS, RECOMMENDATIONS, AND CONCLUSIONS

In this chapter, the reader will find a summary of the study, discussion on the data exploration, discussion of the findings presented by research question, implications for practice, recommendations for further research, and conclusions.

Summary of the Study

Nagy and Herman (1987) trenchantly related that oral language development is a significant factor in vocabulary development, which connects to comprehension, which connects to educational success, which, in turn, is often related to success in life. Specifically, with young children it has been important to realize that vocabulary knowledge in kindergarten and first grade has been a significant predictor of reading comprehension in the middle and secondary grades (Cunningham & Stanovich, 1997; Scarborough, 1998).

A primary need of bilingual and ESL teachers has been to assess and evaluate English acquisition of their students. The assessment that ELL teachers should employ should inform day-to-day instructional decisions, communicate progress to the students and to the parents, identify students in need of additional instruction, and evaluate program effectiveness: in essence teachers must effectively assess language growth (Tinajero & Hurley, 2001).

The purpose of this study was to create and validate a curriculum-based instrument to measure oral proficiency and expressive vocabulary of kindergarten students. The instrument was denominated the Semantic and Syntactic Scoring System (S4) for the STELLA Vocabulary and Oral Proficiency Protocol. A secondary purpose of this study was to compare the performance of students who participated in instruction under two customary ELL programs: Transitional Bilingual Education (TBE) and Structured English Immersion (SEI). This study was conducted in the context of the English and Literacy Acquisition (ELLA) (R305P030032) grant (Lara-Alecio et al., 2003) in which students were provided instruction in four ELL instruction models: TBE control, TBE experimental, SEI control, and SEI experimental.

This chapter includes the findings presented in Chapter IV in terms of extant literature in language test construction, expressive vocabulary, oracy, and curriculum-based assessment as applicable to each research question. The strengths and limitations of this study will be discussed here. Then the implications of this study for theory and praxis are presented.

The discussions that follow are organized according to the research questions that guided this study:

1. To what extent can a curriculum-based measure be developed and validated to measure oral proficiency and vocabulary knowledge among ELLs who are participating in a controlled oral language development intervention?

2. To what extent can teachers use the Semantic and Syntactic Scoring System (S4) for the STELLA vocabulary fluency measure with minimal training to accurately assess students' vocabulary knowledge and oral proficiency?
3. To what extent does the developed curriculum-based assessment instrument, Semantic and Syntactic Scoring System (S4) differentiate the level of knowledge regarding expressive vocabulary and oral proficiency of kindergarten students participating in the STELLA intervention under two different programs: enhanced Traditional Bilingual Education and the enhanced Structured English Immersion Program in comparison to the Revised Woodcock Language Proficiency battery (WLPB-R) (language and vocabulary subtests)?

Discussion by Research Questions

Research Question 1

Research questions 1: To what extent can a curriculum-based assessment instrument be developed and validated to measure oral proficiency and vocabulary knowledge among ELLs who are participating in a story retell intervention? To answer this question validity of the S4 was examined by comparing the S4 to two extant commercial measures on language and vocabulary (subtests of the WLPB-R and ITBS). Second reliability of the S4 was examined through intrarater and interrater reliability.

Validity. The S4 was tested for concurrent validity with other measures: WLPB-R (Picture Vocabulary, Listening Comprehension, and Vocabulary Analogies subtests) and the ITBS (Vocabulary and Word Analysis subtests). The WLPB-R and ITBS subtests were considered the most comparable oral language and expressive measures to the S4. The range of the Pearson correlation coefficients (.133 to .457) for the S4 correlated with the WLPB-R subtests and ITBS subtests were low although statistically significant ($p<.01$). Even when the data were split to evaluate concurrent validity by group membership: TBE-C, TBE-E, SEI-C, and SEI-E, the range of the coefficients expanded but remained low (.025 to .509), but they were also statistically significant ($p<.01$). It is probable that statistical significance was reached because of the large sample size (n= 905) of this study. The S4 correlated higher with the WLPB-R Listening Compression and Picture Vocabulary than any other measure. And it is interesting to note, that commercialized measures subtests did not correlate highly with themselves. Only the WLPB-R Picture Vocabulary and Listening Comprehension (r= .652) were in the *moderate* range.

In referring to Henning's definition on validity (as provided in Chapter IV) Alderson, Clapham, and Wall (1995b, p.170) ascertained that validity is not an all-or-nothing matter and that it is important for test users to use their own (or somebody else's judgment) when deciding if a measure is valid for their particular intended use. It is important to remember that the S4 was a curriculum-based alternative assessment and therefore measured different

things than the standardized-commercial measures on vocabulary and language.

Tinajero and Hurley (2001) reiterated that traditional assessment techniques [such as standardized-commercial norm-referenced tests] are often incongruent with ESL classroom practices. And teachers need to use authentic assessment which are easy to use, economical, an integral part of instruction, account for learning contexts, and which chronicle language growth and development for ELLs (Tinajero and Hurley). In language assessment and when emphasizing classroom-based assessment (rather than standardized, large scale testing), criterion-referenced testing is of prominent interest more so than with norm-referenced testing and if a test can provide instructional value then the distribution of scores along a continuum is of little value when the test provides information on specific objectives (Brown, 2004). A trend has emerged to supplement traditional test designs with alternatives that are more authentic in their elicitation of meaningful communication. Table 31 highlights the differences between traditional and alternative assessment according to Brown (2004).

Table 31

Traditional and Alternative Assessment

Traditional Assessment	Alternative Assessment
One-shot, standardized exams	Continuous long-term assessment
Timed, multiple-choice format	Untimed, free response format
Decontextualized test items	Contextualized communicative tasks
Scores suffice for feedback	Individualized feedback and washback
Norm-referenced scores	Criterion-referenced scores
Focus in the right answer	Open-Ended, Creative Answers
Summative	Formative
Oriented to Product	Oriented to Process
Non-interactive Performance	Interactive Performance
Fosters Extrinsic motivation	Fosters intrinsic motivation

Note: (Brown, 2004) author adapted this from Armstrong (1994) and Bailey (1998)

The S4 adheres to the characteristics of alternative assessment. In order for the S4 to be valid it needs to have enough sameness with measures that purport to measure the same construct (oral proficiency and expressive vocabulary). However, in order for the S4 to offer an authentic curriculum-based alternative to traditional assessment, the S4 should be distinct and thus would not correlate highly with other measures. “Tests can be invalidated by too high correlations with other tests from which they were intended to differ” (Campbell & Fiske, 1959, p.81). Campbell and Fiske (1959) long ago pointed out the

fallacy in assuming that correlations between measures that used the same method to assess the same construct ‘proved’ the validity of a new measure.”

Reliability. The second aspect of the first research question pertained to reliability of the S4, specifically with intrarater and interrater reliability. It is important to establish intra-rater reliability at the end of examiner training or routinely during marking.

> The only way in which intra-rater reliability can be established is by getting examiners to re-mark scripts they have already marked. This will only make sense if the first marks are not on the scripts… and the correlation between the first marks and the second marks, and their respective means and standard deviations can then be checked, and suitable action taken if intra-rater reliability proves to be low. (Alderson et al., 1995b, p. 136)

In developing the S4, there were 7 intrarater check points. Three statistics (Cramer’s V, Kappa, and percent agreement) were calculated each time for intrarater reliability. Initially the intrarater reliability correlations were adequate indicating fair agreement with each rating. Once the S4 was expanded from four to five descriptors then intrarater reliability improved to *moderate* and to *good* agreement. Then, once the S4 was put into flowchart format, the agreement increased to *good* and *very good* agreement. The reason that intrarater reliability statics improved from the first to the seventh intrarater reliability check was that the S4 improved: both in defining the descriptors and in format (flowchart). The flowchart format was attributed with the higher interrater agreements because it functions as a decision-making flowchart, how one answer the first question determines whether one can go on and possibly assign

more points for the sentence that one is rating or whether the highest possible points have already been awarded.

Interrater reliability was tested as part of creating the S4. The S4 underwent several iterations and each time correlation coefficients, effect sizes, and percent agreement improved. During the first iteration, with two ELLA researchers and three graduate students, the descriptors were evaluated and modified according to the feedback provided by them. In the second iteration, the rating of zero points was added for *No Response* or *Response in a language other than English*. The third iteration took place with four ELLA grant coordinators. Their approach to the S4 was a little different because these coordinator had previous knowledge of the target words and how they were taught. Some were inclined (consciously or unconsciously) to rate responses based on the context that the target word was taught. Whereas, those not directly affiliated with the grant accepted multiple meanings for the given words and did not base their scoring on the context in which the word was learned. The grant coordinators, perhaps, more closely resemble the future users of the S4: elementary teachers. So it is important to impart to them that they should accept sentences that reflect different contexts, different from the context the word was taught. Each training and reliability check informed instrument development and clarification. The last interraters checks were conducted when the scale was incorporated into a flow chart to help guide with the scoring. The second, training was with graduate students in bilingual education and a former

elementary classroom teacher. They participated in the S4 training and then rated students' responses, which were used to calculate interrater reliability. During the session if there were ambiguities those were noted and improved for the next training session. The final training and iteration took place with a different group of graduate students and former bilingual and ESL teachers. Again, training for them was refined from the previous ones and it seems that they had higher correlation coefficients (.900 – 1.00), higher effect sizes (.81 – 1.00), higher percent agreement (72%), and higher percent correct (72%, 83%, and rater 78%).

> In principle, a test cannot be valid unless it is reliable. If a test does not measure something consistently, it follows that it cannot always be measuring it accurately. On the other hand, it is quite possible for a test to be reliable but invalid…therefore, although reliability is needed for validity it alone is not sufficient. (Alderson et al., 1995b, p. 187)

Research Question 2

Research question 2 was: *Can teachers use the Semantic and Syntactic Scoring System (S4) for the STELLA Vocabulary Fluency Measure with minimal training to accurately assess student performance?*

It was important to evaluate whether teachers would be able to accurately use the STELLA Vocabulary Fluency protocol and the S4 scoring system with minimal training. Therefore, four elementary, bilingual teachers were asked to review the training manual and score two randomly selected student samples. The first sample was scored for feedback. In the manual, the teachers were able to compare their answers to the correct answers and read the rational behind

each score for each word. Then the second scoring was submitted for interarater re The third iteration took place with ELLA grant coordinators. Their approach to the S4 was a little different because these coordinator had previous knowledge of the target words and how they were taught. Some were inclined (consciously or unconsciously) to rate responses based on the context that the target word was taught. Whereas, those not directly affiliated with the grant accepted multiple meanings for the given words and did not base their scoring on the context in which the word was learned. The grant coordinators, perhaps, more closely resemble the future users of the S4: elementary teachers. So it is important to impart to them that they should accept sentences that reflect different contexts, different from the context the word was taught. The results with this group were much improved. The correlation coefficients were high (.955 to 1.00),and the coefficient of determinations were high (.91 to 1.00). However, in accounting for percent agreement (72%) and percent correct (83% - 100%) the results decreased. Perhaps, if these teachers had been trained and also had use of the materials, based on the patterns in the previous iterations, they might have been able to increase scores on chance agreement (percent agreement) and percent correct..

It is possible that the teachers would have scored even higher had they participated in an in-person training using the S4. There are a couple of concerns with raters that were ameliorated by the flowchart design of the S4. For example, teachers tend to drift away from other raters with whom they use

to agree and they begin to redefine the rating rubric for themselves (Nitko & Brookhart, 2007). Second raters tend to engage in reliability decay, which means that the rater applied the rubric correctly but then with the passing of time, the ratings become less consistent, across students and across raters (Nitko & Brookhart). The scores of the teacher raters were less subject to these issues because they had a flowchart-format for the S4, which created an inherent consistency in the rating process.

It was important to evaluate whether the others could use the materials and score consistently and accurately without necessitating an in-person training for each individual that would employ this instrument.

Research Question 3

Research question 3 was: *To what extent does the developed curriculum-based assessment instrument, Semantic and Syntactic Scoring System (S4) differentiate the level of knowledge regarding expressive vocabulary and oral proficiency of kindergarten students participating in the STELLA intervention under two different programs: enhanced Transitional Bilingual Education and the enhanced Structured English Immersion Program in comparison to the Revised Woodcock Language Proficiency Battery (WLPB-R) (language and vocabulary subsets)?*

The students in the enhanced treatments participated in the STELLA intervention and other intervention components. The impact of STELLA and the other interventions could not be disentangled to qualify any statements that

STELLA was the intervention that most contributed to oral proficiency and vocabulary knowledge of the students, in this particular study. All that can be stated is that there were strong intervention (control versus experimental) effects for the S4 and there were strong interaction effects (program type and intervention). No causal statements can be made that the STELLA intervention was the only intervention making the difference for the groups' performance in the standardized and commercial tests used in the ELLA project. However, since the Project STELLA Vocabulary Fluency Measure and the S4 are specifically designed to test the STELLA curriculum the performance on that measure is more indicative of the effectiveness of the STELLA intervention component. It is true that the other intervention components can influence oral proficiency and word knowledge overall, so the above is not a definitive correlational statement. It is just clear that because a curriculum-based test is being used to test the intervention, it is a better measure of that intervention than other extant, standardized, and commercial measures.

Table 32 provides the Summary of the Type III Effects of the S4. And it was evident that the program type (bilingual versus SEI) was not statistically significant ($p<.05$). The intervention (comparing the performance of the control groups to the performance of the experimental [enhanced] groups in TBE and SEI]) was statistically significant ($p>.05$). Furthermore, the interaction of the program type and intervention was not statistically significant ($p<.05$).

Table 32

Summary of the Type III Effects of the S4

	S4	PV	LC	VA
Program Type (TBE/SEI)	No	Yes	Yes	No
Intervention (Ctrl/Exp)	Yes	No	No	No
Interaction	No	Yes	No	No

In looking at the scores that the students obtained on the S4 (see Chapter IV, Figure 3) it is important to consider the stages of language growth for English Language Learners (ELLs). Just as individuals acquire their first language, there are sequences of stages that are evident in the second language acquisition process, too (Tinajero & Hurley, 2001). The second language acquisition process begins with a silent period or preproduction stage, during this stage children are listening and assimilating the sounds and structures of the language. Since the students in this study are kindergarten ELLs it is to be expected that many of them would have been at the preproduction stage. And this perhaps, explains why many of the scores were 0 (no response given) or 18 (merely repeating of the target word). Tinajero and Schifini (1997) also relate that students undergo growth language spurts. It can be customary to see a surge in vocabulary knowledge but a lack in grammatical

ability to control and use that new vocabulary. The S4 takes this into consideration by measuring the sentence in light of both semantics and syntax and not just one or the other. Because to focus on one or the other provides limited information, since expressive vocabulary is confounded in grammar/syntax and visa versa. Taking both of these into account is what the S4 looks at as oral proficiency.

During Project ELLA a study was conducted to ascertain language use and communication modes used among the four classroom designations of interest: TBE control, TBE experimental, SEI control, and SEI experimental. That project ELLA study collected data using the instrument, Transitional Bilingual Observation Protocol (TBOP) (Lara-Alecio & Parker, 1994). The study found that communication in the classrooms was different. In the experimental classrooms aural-verbal modes of communication were used to a greater extent (97%) as opposed to in the typical classrooms where the aural-verbal mode of communication was less (70 %). Additionally, English was used at a higher rate during ESL instruction segments in the experimental designation as opposed to the control designations.

Other studies conducted with this data support that the TBE control group underperforms when compared to the other groups(TBE experimental, SEI control, and SEI experimental). The Tong (2006) study found that there was a statistically significant difference between initial levels of oral English proficiency (as measures by the WLPB-R) between the SEI control and experimental

groups. The experimental groups seemed to have a lower level of English language proficiency at the onset and yet was able to demonstrate higher rates of English acquisition, for both TBE and SEI. This is consistent with studies conducted under Project ELLA. The Tong (2006) study concluded that the starting level in oral proficiency does not matter as much as the language of instruction when it comes to the development of oral English as a second language.

Another study by Quiros (2008) also found that there was a statistical difference in performance when the control groups were compared to the experimental for TBE and SEI, respectively, across kindergarten, first-, and second-grades. The students in the enhanced treatments of TBE and SEI were receiving instruction in STELLA and data were collected to determine how they performed on a measure of Story retell in comparison to the control groups for both TBE and SEI.

The study by Tong, Lara-Alecio, Irby, Mathes, and Kwok (2008b) reported that students in the TBE experimental had a statistically significant improvement over TBE control group when it came to scores that reflect listening comprehension performance across several grade-levels.

> Indeed, educational interventions frequently yield “fan” spreads reflecting differential impacts for students starting at different levels…less able students over the course of an intervention may stay about the same or slightly improve. More able students may not only improve, but may even improve more drastically than their less able counterparts. This dynamic reflects the fact that pretest achievements scores involve estimated abilities at a given point in time, but may involve as well differential rates of learning. (Thompson, 2006, p. 56)

This study corroborates the above mentioned study because it is conducted with the first year of data, kindergarten scores for the 4 groups. One would expect that with time (subsequent grade-levels) the students in the enhanced treatments would continually outperform those of the control groups. And one would expect that those in the SEI could outperform those in the TBE when measuring English proficiency, since SEI affords their instruction exclusively in English. However, this study looked at the scores in kindergarten before educational impacts of “fan” spread considerations.

Limitations

The present study provided a concerted effort to add to the limited body of knowledge on oral language proficiency assessment for young ELLs. However, there were some limitations to the study. First, the mixed model analysis employed in this study was robust and would have allowed for some generalizations to be made beyond the participants and parameters of this study; but it is important to reiterate that the data for this study were from a federally funded longitudinal grant and because of the nature and context of the study the design was quasi-experimental, thus diminishing generalizability. “Only experimental designs allow us to make definitive statements about causality, although other research designs may suggest the possibility of causal effects” (Thompson, 2006, p. 24).

The data for the investigation were archival and extracted from a larger study; consequently, covariates were used to account for the lack of pretests

scores in the Project STELLA Vocabulary Fluency Measure. It would have been preferable to have actual pretests on the Project STELLA Vocabulary Fluency Measure and to have scored them using the S4. A pretest and posttest score would have permitted a more direct comparison. Since there was no pretest scores on the S4 it had to be assumed that the students had little prior knowledge of the target vocabulary words prior to receiving instruction and being tests on the words.

In addition, the study only analyzed the performance of kindergarten students on the Project STELLA Vocabulary Fluency Measure using the S4. It is possible that the age of the participants affords differential information using the S4 as opposed to using the S4 with older students. For example, in the Snow et al (1987) study it became evident that the strength of the correlation of definitions and quality of definitions increased over grades – so age became a factor in student performance. This study was developed in light of kindergarten participants. It is possible that the S4 instrument will not provide sufficient information or distinction among students as they become more proficient as they get older. It is possible that the scores on S4 will cluster to the higher end as the students in this study or other students are tested at higher grade-levels. When ratings cluster so that it is not possible to distinguish a student's performance from other students then the scores become unreliable and the validity of the scores are also reduced (Nitko & Brookhart, 2007, pp. 281-282).

Another inherent limitation of this study exists in the statement made by Fulcher (1997); he argued that speaking tests were particularly problematic when considering reliability, validity, practicality, and generalizability.

For oral administration of tests, research has not addressed how much time should elapse between questions and responses and this could potentially affect results (Murphy, 1997) and when administering the Project STELLA Vocabulary Oral Proficiency Protocol the time was not always consistent in how long the examiner would pause before they decided to move on the next word, if the student delayed in responding was delayed. The testing protocol required that the examiner wait one minute before assuming that the student was not going to respond. However, a few of the examiners did not wait one minute for the student to provide a response. This could have affected the scores obtained. In addition, different elicitation tasks and test methods influence results differently, limiting interpretation of constructs (Bachman, 1990).

External validity is also limited in this study. This study compares the SEI and TBE in some of the same schools and this constrains the generalizability of the results from these comparisons because of the small number of eligible schools and because I was dealing with just one district, so district and school effects could not be separated.

The statistical significance obtained by the analyses in this study, should be qualified with the following consideration: very small and unimportant effects can turn out to be statistically significant because of a large N (Field & Hole,

2003). The *n* for this study is considered a large. However, it is important to note that Cohen (1990) pointed out that a non-significant result should not be interpreted as meaning that there is no difference between the means or that there is no relationship between the variables. Also, Cohen points out that the null hypothesis is never true because it never is, a big *n* will always make small differences significant. Given that the analyses in this study were selected for being robust, that ameliorates the concern with inflated statistical significance.

Finally, it is important to note that any measurement, even physical measurements, will invariably generate scores that include some degree of measurement error or some degree of unreliability (Thompson, 2003).

Implications for Practice

In their synthesis Saunders and O'Brien (Saunders & O'Brien, 2006) found few 2L studies that focused on specific aspects of LA such as vocabulary, specific grammatical forms, or pragmatic patterns. Although the opposite is true in L1 Acquisition. Thus, there is little evidence about L1 development and little empirical basis in which to base interventions that promote specific language development. Therefore, this study adds to the theory literature corpus and the praxis corpus on oral language.

Utility

Research has indicated that classroom-based assessment has potential for accurately ascertaining student knowledge and competence (Airasian, 1991; Shepard, 1995; Stiggins, 1999). However, as perceived by O'Neil (O'Neil, 1992),

most classroom-based assessment methods tended to be informal and teachers needed increased expertise in this type of assessment. Teachers should design instructional modifications based on assessment data in order to help students improve (Frey & Hiebert, 2003). However, it has been unusual for teachers to do that because most teachers do not make inferences or interpret data for planning instructional interventions (Butler & McMunn, 2006).

The scores obtained using the S4 on an instrument such as the Project STELLA Vocabulary Fluency Measure can be used to contribute to program placements. It can also assist teachers in creating cooperative groups based on language ability. Teachers could also rank order the scores and with that information create homogenous or heterogeneous cooperating learning groups or student pairings. The S4 allows teachers to conduct long-term assessment on students. Long-term and continuous assessment give educators better insight into students' understanding and knowledge (Hurley & Blake, 2001).

Schrank et al (1996) stated that oral proficiency tests are often used to determine ELL program placement (establish or deny eligibility for instruction in English or another language) and the caveat is made that when using tests of oral proficiency they should surpass mere measurement of Basic Interpersonal Communications Skills (BICS). It is important that these tests include Cognitive Academic Language Proficiency (CALPS) (Ibid). The Project STELLA Vocabulary Fluency Measure and the S4 can be used as an instrument that covers both dimensions. It does allow for conversational skills to become

evident through the use of oral sentence production; however, it extends this further by rating the sentences based on semantics, syntax, and word knowledge. More validation is needed to determine if the S4 correlates sufficiently with other measures of BICS and CALPS that are customarily used by school districts.

Cost

Educators and advocates have begun arguing for educational reform that would de-emphasize standardized and large-scale tests in favor of structuring budgets to accommodate the use of contextualized, communicative performance-based assessment which inform curricula (Brown, 2004). The S4 does not cost much to replicate. It can be used as frequently or infrequently as wanted.

Recommendations for S4

This study presents an initial attempt at creating a curriculum-based instrument to measure English oral proficiency and expressive vocabulary. Since, "…problems with a test or associated procedures may only emerge once the test has been in operation for some time" (Alderson et al., 1995b, p. 218), further research should continue to examine additional psychometric properties of the S4 or similar measures because the instrument is mostly supported by theory and would require more supportive validation evidence. The instrument needs more validation evidence both in looking at the theory behind the descriptors and the psychometric properties of the instrument.

Validity

On consideration in addressing validity if the S4 has to do with the development of children in terms of language. This instrument was developed and tested with kindergarten students, only. It is possible the S4 may need to be modified to provide greater validity for older children.

> Young children very reasonably respond to a question like 'What's a hat?' with 'you wear it, and such a response is tolerated if the child is young enough. Older children, on the other hand, are expected to respond to such questions by giving 'formal definitions,' which conform to particular standards for form as well as for content, for example. 'A hat is an article of clothing worn on the head.' (Snow, Cancino, de Temple, & Schley, 1991, p. 90)

Young children include incidental, highly personal, and idiosyncratic information when providing definitions (Snow et al., 1991). This was evident in that most of the sentences produced by the students were personal, involving themselves or their family in the sentence. Perhaps, this is a characteristic that can also be measured to further distinguish language levels and provide validity for the instrument in upper grades. Henning defined validity as:

> Validity in general refers to the appropriateness of a given test or any of its component parts as a measure of what it is purported to measure. A test is said to be valid to the extent that it measures what it is suppose to measure. It follows that the term *valid* when used to describe a test should usually be accompanied by the preposition *for*. And test then may be valid for some purposes, but not for others. (Henning, 1987, p.89)

There are different aspects of validity such as content/rational validity, concurrent/empirical validity, construct validity, and the newer criterion validity (Alderson et al., 1995b). All of these validities can be further explored.

Another limitation could be with score stability. It would be good to test the S4 across some time sample with different words and see if there is stability despite the target words changing and despite the expect growth with time.

Alderson et al (1995a, p. 185) presented three points that are to be considered concerning measuring oral language proficiency. The results on the S4 could be compared to teacher ranking of student oral language ability because teachers usually have a fair idea of the levels of ability of the students in his or her class and this comparison would allow for further validating the content of the instrument. Also construct validity could be examined by providing the instrument to experts and seeing if they find that the instrument measures the construct of oral language as defined and intended in the study (Alderson). Third the scores on the Project STELLA Vocabulary Fluency Measure using the S4 could be tested in terms of biodata of students (gender, age, first language, number of years studying the language) and one would expect that they there would be difference according to the biodata.

Reliability

"In practice, second language learning studies yield correlations whose maximum values rarely approach +1 and are more likely to be in the order of 0.30 – 0.60." (Skehan, 1989, p. 13) .If statistical significance is attained, as was in most of the interrater reliability, means that the results were unlikely to have arisen by chance. In practice, for correlation coefficients, establishing significance is dependent on two factors – the magnitude of the relationship

found, and the sample size: the larger the sample size the lower the correlation coefficient needed to claim significance.

> The problem for most language testers is that in order to maximize reliability it is often necessary to reduce validity. Some people would argue that reliability must be sacrificed to achieve validity. Yet we cannot have validity without reliability. In practice, neither reliability nor validity are absolutes: there are degrees of both, and it is commonplace to speak of a trade-off between the two, you maximize one at the expense of the other. Which you choose to maximize will depend on the test's purpose and the consequences for individuals of gaining an inaccurate result. (Alderson et al., 1995b, p. 187)

Intrarater Reliability

Intrarater reliability recommendations: A solution to intrarater reliability unreliability is to read through about half of the tests before rendering final scores and the recycling back through the whole set of tests to ensure even handed-judging (Brown, 2004). For example, in this study the primary rater for the data was the researcher. However, if other raters are going to rate for a research study or classroom use then additional interrater issues should be addressed. To improve intrarater reliability teachers could do the following with their class set of tests to improve their intra-rater reliability. "Also, the instrument needs to be evaluated for effectiveness in actual classroom settings with teachers using the instrument. Some of the things that could be evaluated with teacher use to inform reliability: leniency error, severity error, central tendency, Halo effect, personal bias, logical errors, rater drift, and reliability decay (see Nitko & Brookhart, 2007)

Interrater Reliability

Recommendations for interrater reliability is the use of routine double-marking, in which every exam is scored by two examiners and these two are averaged. Before this an administrator could compare them and if the scores are similar then they can be averaged. Similar means that they are less than two points apart, very different means that they are two points or more apart in a five-point-rating scale, if this is the case then the raters need to study the rating scale again (Alderson et al., 1995b) They recommend double marking because it allows some variations because in language testing differences of opinion between examiners could be legitimate.

Concluding Remarks

"The statistics are clear-ELLs will constitute an ever-expanding and, thus, important portion of the school-age population. Effective education for ELLs means planning for their and the nation's future" (Genesee et al., 2006, p. 233).

> Having now entered the 21st century, children in elementary schools today will need an unprecedented level of oracy to meet the challenges of the new century. The revisions required may not lend themselves to packaged programs or written materials. Oral language skills and concepts are best developed in situations that imitate life. Constructing such learning experiences will not be easy and will require extensive study, and development on the part of teachers. With that in mind, it appears that an immediate start is warranted. (Pinnell & Jaggar, 2003, p.904)

Through my study, I created an instrument that meets the need to assessing oral proficiency and vocabulary knowledge of young ELLs. Perhaps, my study will meet the identified need of helping teachers acquire detailed

descriptions of their students spoken language skills, which is most important when teachers work with students from diverse populations (Riley & Burrell, 2007). The focus on oracy is important because Hiebert, Pearson, Taylor, Richardson, and Paris emphasized that “oral language is the foundation on which reading is built, and it continues to serve this role as children develop as readers (1988).

Furthermore, the S4 was created in adherence to principles of quality assessment as defined by McMillan (2007). McMillan defined quality assessment as assessment that adheres to specific psychometric standards, validity and reliability, among other principles. Accordingly for teachers, the measure of quality for a test exceeds psychometric soundness and requires that the test assess what students can do based on the curriculum with the intent of informing instruction. An expanded definition of quality assessment had the following criteria: (a) clear and appropriate learning targets, (b) appropriateness of assessment methods, (c) validity, (d) reliability, (e) fairness, (f) positive consequences, (g) alignment, and (h) practicality and efficiency (McMillan, 2007, p.57). Brown (2004, p.19) called attention to the attributes of effective tests by saying that they should be practical tests and not excessively expensive, stay within appropriate administration time constraints, be relatively easy to administer, and have a scoring/evaluation procedure that is specific and time-efficient. Tinajero and Hurley (2001, p.35) outlined three specific purposes for authentic assessment: (a) the measures need to be an integral part of

instruction, (b) the measures need to consider the learning context of the individual child, whether they are working alone or with others, and (c) assessments must provide insight into the development and growth of language and academics. "…we have a very limited understanding of specific aspects of L2 oral language development and, thus, little empirical basis for planning educational interventions that would promote language development in specific ways." (Saunders & O'Brien, 2006, p.15) Because teachers need to continuously evaluate the strengths and weaknesses of their students and adjust their teaching in order to meet the language and literacy needs of the students (Fillmore & Snow, 2000) and classroom-based assessment helps teachers identify instructional needs and modify instruction (Hurley & Tinajero, 2001) to continue using this instrument and fine-tuning it should prove fruitful in enhancing ELL instruction.

As Loban (1976, p.90) trenchantly stated, "Complex truth is always an aggregate; each of it offers only part of an evolving mosaic." I have attempted to contribute to the mosaic of vocabulary acquisition and oral proficiency with a study that begins to look at theoretical basis and assessment of oral proficiency coupled with expressive vocabulary knowledge in young ELLs.

REFERENCES

Acock, A. C., & Gordon, R. S. (1979). A measure of association for nonparametric statistics. *Social Forces, 57*(4), 1381-1386.

Adamson, H. D. (2005). *Language minority students in American schools*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Airasian, P. W. (1991). Perspectives on measurement instruction. *Educational Measurement: Issues and Practice, 10*(1), 13-16.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Testing, 35*, 79-113.

Alderson, J. C., Clapham, C., & Wall, D. (1995a). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995b). *Language testing construction and evaluation*. Cambridge: Cambridge University Press.

Allison, P. D. (1999). *Multiple regression*. Thousand Oaks, CA: Pine Forge Press.

Altman, D. G. (1991). *Practical statistics for medical research* (1st ed.). London: Chapman and Hall.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anglin, J. M. (1993). *Vocabulary development a morphological analysis*. Chicago, IL: The University of Chicago Press.

Armstrong, T. (1994). *Multiple intelligences in the classroom*. Philadelphia: Association for Curriculum Development.

August, D., Carlo, M., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research and Practice 20*(1), 50-57.

August, D., & Shanahan, T. (2006). *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers: Center for Applied Linguistics.

Bachman, L. F. (1988). Language testing - SLA research interfaces. *Journal of Applied Linguistics, 9*, 193-209.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F., & Clark, J. L. D. (1987). The measurement of foreign/second language proficiency. *Annals of the American Academy of Political and Social Science, 490*, 20-33.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449-465.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* Oxford: Oxford University Press.

Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions, and directions*. Cambridge, MA: Heinle and Heinle.

Baker, C. (1995). Bilingual education and assessment. In B. M. Jones & P. Ghuman (Eds.), *Bilingualism, education and identity*. Cardiff, Wales: University of Wales Press.

Baker, C. (1996). *Foundations of bilingual education and bilingualism* (2nd ed.). Bristol, PA: Multilingual Matters, Ltd.

Baker, K. (1992). Ramírez et al.: Misled by bad theory. *Bilingual Research Journal, 16*(1-2), 63-89.

Ballard, W. S., Tighe, P. L., & Dalton, E. F. (1980). *IDEA Oral Language Proficiency Test 1*. Bream, CA: Ballard & Tighe.

Barnhart, J. E. (1990). Differences in story retelling behaviors and their relation to reading comprehension in second graders. In J. Zutell & S. McCormick (Eds.), *Literacy theory and research: Analyses from multiple paradigms*. Chicago, IL: The National Reading Conference, Inc.

Bear, D. R., & Helman, L. (2004). Word study for vocabulary development in the early stages of literacy learning. In J. F. Baumann & E. J. Kame'enui (Eds.), *Vocabulary instruction: Research to practice*. New York: Guildford Press.

Beck, I. L., & McKeown, M. G. (1987). Getting the most from basal reading selections. *Elementary School Journal, 87*, 343-356.

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York: Guilford Press.

Berry, K. J., & Mielke, P. W. J. (1988). A generalization of Cohen's Kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement, 48*, 921-934.

Biemiller, A. (2003). Oral comprehension sets the ceiling on reading comprehension. *American Educator, 27*(1), 23.

Biemiller, A. (2004). Teaching vocabulary in the primary grades: Vocabulary instruction needed. In J. F. Baumann & E. J. Kameenui (Eds.), *Vocabulary instruction: Research to practice.* New York:The Guildford Press.

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology, 93*(3), 498-520.

Blank, M., & Frank, S. M. (1971). Story recall in kindergarten children: Effect of method of presentation on psycholinguistic performance. *Child Development* (42), 299-312.

Bloom, B. S., Madus, G. F., & Hastings, J. T. (1981). *Evaluation to improve learning*. New York: McGraw-Hill.

Bloom, P. (2000). *How children learn the meaning of words*. Cambridge, MA: The MIT Press.

Bock, D. G., & Bock, E. H. (1984). The effects of positional stress and receiver apprehension on leniency errors in speech evaluation: A test of the rating error paradigm. *Communication Education, 33*, 337-341.

Boehmm, A. E. (1992). Glossary of assessment terms. In L. R. Williams & D. P. Fromberg (Eds.), *Encyclopedia of early childhood education*. New York: Garland.

Bozdogan, H. (1987). Model selection and Akaike's Information Criteria (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Bracey, G. (1989). The $150 million redundancy. *Phi Delta Kappan, 70,* 698-702.

Breunig, N. A. (1998). *Measuring the instructional use of Spanish and English in elementary transitional bilingual classrooms.* Unpublished doctoral dissertation, Texas A&M University, College Station.

Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112 - 140). Cambridge University Press.

Britton, J. (1970). *Language and learning*. Coral Gables, FL: University of Miami Press.

Brooks, K. (1957). The construction and testing of forced choice scale for measuring speaker achievement. *Speech Monographs, 24*, 65-73.

Brown, A. (1975). Recognition, reconstruction and recall of narrative sequences of preoperational children. *Child Development, 46*, 155-156.

Brown, H. D. (2004). *Language assessment principles and classroom practices*. White Plains, NY: Pearson Education, Inc.

Bryk, A., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Butler, S. M., & McMunn, N. D. (2006). *A teacher's guide to classroom assessment: Understanding and using assessment to improve student learning*. San Francisco, CA: Jossey-Bass.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 89-112.

Carlisle, J. F. (1988). Knowledge of derivational morphology and spelling ability in fourth, sixth and eighth graders. *Applied Psycholinguistics, 9*, 247-266.

Carlo, M., August, D., McLaughlin, B., Snow, C., Dressler, C., Lippman, D., et al. (2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39*(2), 188-215.

Catts, H., Fey, M., Tomblin, J. B., & Zhang, X. (2002). A longitudinal investigation of reading outcomes in children with language impairments. *Journal of Speech, Language, and Hearing Research, 45*(6), 1142-1157.

Chalhoub-Deville, M. (1996). Performance assessment and the components of the oral construct across different tests and rater groups. In M. Milahovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC) Cambridge and Arnhem*. Cambridge: Cambridge University Press.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of theory of syntax*. Cambridge, MA: MIT Press.

Clark, J. L. D. (1975). Theoretical and technical considerations in oral proficiency testing. In R. L. Jones & B. Spolsky (Eds.), *Testing language proficiency*. Arlington, VA: Center for Applied Linguistics.

Close, R. A. (1982). *English as a foreign language*. London: George Allen and Unwin.

Coady, J. (1997). L2 vocabulary acquisition through extensive reading. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition*. New York: Cambridge University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cohen, J. (1990). Things I have learned so far. *American Psychologist, 45*, 1304-1312.

Cole, D. A., S.E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among dependent variables. *Psychological Bulletin, 115*(3), 465-474.

Collier, V. P. (1992). A synthesis of studies examining long-term language minority student data on academic achievement. *The Bilingual Research Journal, 16*(1-2), 187-212.

Crawford, J. (2000). *At war with diversity: U.S. language policy in an age of anxiety*. Tonawanda, NY: Multilingual Matters.

Cronbach, L. J. (1971). Test Validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California Department of Education (Ed.), *Schooling and language minority students: A theoretical framework* (pp. 3-49). Los Angeles, CA: California State: Evaluation, Dissemination, and Assessment Center.

Cummins, J. (1986). Empowering minority students: A framework for intervention. *Harvard Educational Review, 56*(1), 18-35.

Cummins, J. (1996). *Negotiating identities: Education for empowerment in a diverse society*. Los Angeles, CA: California Association for Bilingual Education.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relationship to reading experience and ability 10 years later. *Developmental Psychology*(33), 934-935.

De Avila, E. A., & Duncan, S. E. (1991). *Language assessment scales*. San Rafael, CA: Linguametrics.

Demidenko, E. (2004). *Mixed models: Theory and applications*. Hoboken, NJ: John Wiley & Sons, Inc.

Deno, S. L. (1985). Curriculum-based assessment: The emerging alternative. *Exceptional Children 52*, 219-232.

Dockrell, J., & Messer, D. (2004). Lexical acquisition in the early years. In R. Berman (Ed.), *Language development across childhood and adolescence* (pp. 35-52). Amsterdam: John Benjamins.

Duchan, J. F. (2004). *Frame work in language and literacy*. New York: The Guildford Press.

Dulay, H. C., & Burt, M. K. (1973). Should we teach children syntax? *Language Learning, 23*(2), 245-258.

Dunn, L. M., Dunn, L. M., Robertson, G. J., & Eisenberg, J. L. (1981). *The Peabody Picture Vocabulary Test - Revised (PPVT-R)*. Circle Pines, MN: American Guidance Service.

Dunn, L. M., Dunn, L. M., Williams, K. T., & Wang, J. J. (1997). *Peabody Picture Vocabulary Test III.* Circle Pines, MN: American Guidance Service.

Durgunoglu, A., Nagy, W. E., & Hancin-Bhatt, B. J. (1993). Cross-language transfer of phonological awareness. *Journal of Educational Psychology, 85*(3), 453-465.

Elford, G. W. (2002). *Beyond standardized testing: Better information for school accountability and management*. Lanham, MD: Scarecrow Press, Inc.

Eller, R. G., Pappas, C. C., & Brown, E. (1988). The lexical development of kindergarteners: Learning from written context. *Journal of Reading Behavior, 20*(1), 5-24.

Elley, W. (1991). Acquiring literacy in a second language: The effect of book-based programs. *Language Learning, 41*, 375-411.

Ellis, R. (1985). *Understanding second language acquisition*. Oxford: Oxford University Press.

Ellis, R. (1994). Factors in the incidental acquisition of second language vocabulary for oral input: A review essay. *Applied Language and Learning, 5*(1), 1-32.

Engber, C. (1987). Proceedings of the symposium on the evaluation of foreign language proficiency: Summary of discussion. In A. Valdman (Ed.), *Symposium on the Evaluation of Foreign Language Proficiency* (pp. 1-

14). Bloomington, IN: Indiana University, Committee for Research and Development in Language Instruction.

Field, A. D. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage.

Field, A. D., & Hole, G. J. (2003). *How to design and report experiments*. London: Sage.

Fillmore, L. W., & Snow, C. E. (2000). *What teachers need to know about language. (Contract No. ED-99-CO-0008)*. Washington, DC: Department of Education's Office of Educational Research and Improvement.

Firth, J. R. (1957). *Papers in linguistics*. London: Oxford University Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. Somerset, NJ: John Wiley & Sons, Inc.

Frey, N., & Hiebert, E. H. (2003). Teacher-based assessment of literacy learning. In J. Flood, D. Lapp, J. R. Squire & J. M. Jensen (Eds.), *Handbook of research on teaching the English language arts* (2nd ed.) (pp. 608-618). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Fuchs, L., Fuchs, D., Hosp, M., & Jenkins, J. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*, 239-256.

Fulcher, G. (1997). The testing of L2 speaking. In C. Clapham & D. J. Corson (Eds.), *Language testing and assessment* (Vol. 7, pp. 75-85). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York: Routledge.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal, 64*(4), 428-433.

García-Vázquez, Vázquez, L., López, I., & Ward, W. (1997). Language proficiency and academic success: Relationships between proficiency in two languages and achievement among Mexican-American students. *Bilingual Research Journal 21*(4), 395-408.

García, G. E. (1991). Factors influencing the English reading test performance of Spanish-speaking Hispanic students. *Reading Research Quarterly, 26*, 371-392.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28*(2), 241-265.

Genesee, F., & Hamayan, E. V. (1991). Classroom-based assessment. In E. V. Hamayan & J. S. Damico (Eds.), *Limiting bias in the assessment of bilingual students* (pp. 212-239). Austin, TX: Pro-ED.

Genesee, F., Lindholm-Leary, K., Saunders, W. M., & Christian, D. (2004). English language learners in U.S. Schools: An overview of research findings. *Journal of Education for Students Placed at Risk, 10*(4), 363-385.

Genesee, F., Lindholm-Leary, K., Saunders, W. M., & Christian, D. (Eds.). (2006). *Educating English language learners: A synthesis of research evidence.* New York: Cambridge University Press.

Genesee, F., & Nicoladis, E. (2007). Bilingual first language acquisition. In E. Hoff & M. Shatz (Eds.), *Blackwell handbook of language development* (pp. 324-342). Malden, MA: Blackwell Publishing.

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education*. New York: Cambridge University Press.

Gersten, R., & Baker, S. (2000). The professional knowledge base on instructional practices that support cognitive growth for English-language learners. In R. Gersten, E. Schiller & S. Vaughn (Eds.), *Contemporary special education research: Syntheses of knowledge base on critical instructional issues* (pp. 31-79). Mahwah, NJ: Lawrence Erlbaum Associates.

Gilmore, A. M. (1998). *School entry assessment: The first national picture*. Wellington: New Zealand Ministry of Education.

Good, R. H., & Kaminski, R. A. (2002). *DIBELS oral reading fluency passages for first through third grades (Technical Report No. 10)*. Eugene, OR: University of Oregon.

Good, R. H., Kaminski, R. A., Moats, L. C., Laimon, D., Smith, S., & Dill, S. (2003). *Dynamic indicators of basic early literacy skills* (DIBELS), 6th ed. Longmont, CO: Sopris West.

Goodman, G. S., & Carey, K. T. (2004). *Ubiquitous assessment: Evaluation techniques for the new millennium*. New York: Peter Lang Publishing.

Grosjean, F. (1985). The bilingual as a competent but specific speaker-hearer. *Journal of Multilingual and Multicultural Development, 6*(6), 467-477.

Gullo, D. F. (1994). *Understanding assessment and evaluation in early childhood education*. New York: Teachers College Press.

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

Hadden, B. L. (1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning, 41*(1), 1-24.

Halliday, M. A. K. (1985). *An introduction to functional grammar*. London: Edward Arnold.

Hamayan, E. V. (1995). Approaches to alternative assessment. *Annual Review of Applied Linguistics, 15*, 212-226.

Hargett, G. R. (1998). *Assessment in ESL & bilingual education: A hot topics paper*. Retrieved November 4, 2006 from http://www.nwrac.org/pub/hot/assessment.html

Hargis, C. H. (1995). *Curriculum based assessment: A primer* (2nd ed.). Springfield, IL: Charles C. Thomas.

Harley, B., Allen, P., Cummins, J., & Swain, M. (1990). *The development of second language proficiency*. Cambridge: Cambridge University Press.

Hawkins, R. (2001). *Second language syntax*. Malden, MA: Blackwell Publishers Inc.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics, 32*(4), 341-370.

Henning, G. (1987). *A guide to language testing: Development, evaluation, and research* (4th ed.). Cambridge, MA: Newberry House Publishers.

Hiebert, E. H., & Kamil, M. L. (2005). *Teaching and learning vocabulary: bringing research to practice*. Mahwah, NJ: L. Erlbaum Associates.

Hiebert, E. H., Pearson, B. Z., Taylor, D. M., Richardson, V., & Paris, S. G. (1988). *Every child a reader: Applying reading research in the classroom*. Ann Arbor, MI: University of Michigan: Center for the Improvement of Early Reading Achievement (CIERA).

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (5th ed.). Boston: Houghton Mifflin Company.

Hock, I. (2003). *Test construction and validation*. Budapest, Hungary: Akadémiai Kiadó.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., & Dunbar, S. B. (2006). *Iowa Test of Basic Skills, Forms A, B, C*. Retrieved July 10, 2007, from http://www.riverpub.com/products/itbs/details.html

Howard, E. R., Christian, D., & Genesee, F. (2004). *The development of bilingualism and biliteracy from grade 3 to 5: A summary of findings from the CAL/CREDE study of two-way immersion education*. Santa Cruz, CA: Center for Research on Education, Diversity & Excellence (CREDE).

Hurley, S. R., & Blake, S. (2001). Assessment in the content areas for students acquiring English. In S. R. Hurley & J. V. Tinajero (Eds.), *Literacy assessment of second language learners* (Ch. 5). Needham Heights, MA: Allyn & Bacon.

Hurley, S. R., & Tinajero, J. V. (2001). *Literacy assessment of second language learners*. Boston, MA: Allen and Bacon.

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika, 76*, 297-307.

Hymes, D. H. (1972). Foundations in Sociolinguistics. In J. P. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth, London: Penguin.

Irby, B., Lara-Alecio, R., Quiros, A. M., Mathes, P. G., & Rodriguez, L. (2004). [STELLA Vocabulary Fluency Measure]. Unpublished curriculum-based assessment. Huntsville, TX: Sam Houston State University, Department of Educational Leadership and Counseling.

Irby, B., Quiros, A. M., Lara-Alecio, R., Rodríguez, L., & Mathes, P. G. (2008). What administrators should know about a research-based oral language

development intervention for English language learners: A description of story retelling and higher order thinking for English language and literacy acquisition - STELLA. *International Journal of Educational Leadership Preparation, 3*(2), Retrieved November 29, 2008 from http://ijelp.expressacademic.org

Irby, B. J., Lara-Alecio, R., Quiros, A. M., Mathes, P. G., & Rodriguez, L. (2004). *English language acquisition evaluation research program (Project ELLA): Second annual evaluation report*. Washington, DC: Institute for Educational Sciences, U.S. Department of Education.

Justice, L. M. (2002). Word exposure conditions and preschoolers' novel word learning during shared storybook reading. *Reading Psychology* (23), 87-106.

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 & Language Instruction Education Programs.

King, J. (2004). Calculating a generalized Kappa statistic for use with multiple raters [Microsoft Excel Spreadsheet]. Retrieved from http://www.ccitonline.org/jking/homepage/kappa.xls

King, J. E. (2004). *Software solutions for obtaining a Kappa-type statistic for use with multiple raters.* Dallas, TX: Southwest Educational Research Association.

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *Modern Language Journal, 73*(4), 440-464.

Kreft, I. G. G., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Lara-Alecio, R., Irby, B., & Mathes, P. G. (2006). *Project ELLA: English language and literacy acquisition.* Paper presented at the American Educational Research Association, San Francisco, CA.

Lara-Alecio, R., Irby, B. J., & Mathes, P. G. (2003). *English language and literacy acquisition (Project ELLA)*. U.S. Department of Education, Washington, DC 20202, Contract No R305P030032.

Lara-Alecio, R., & Parker, R. (1994). A pedagogical model for transitional English bilingual classrooms. *Bilingual Research Journal, 18*(3&4), 119-133.

Laufer, B. (1997). The lexical plight in second language reading. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 20-34*).* New York: Cambridge University Press.

Leung, C. B., & Pikulski, J. J. (1990). Incidental learning of word meanings by kindergarten and first grade children thorough repeated read aloud events. In J. Zutell & S. McCormick (Eds.), *Literacy theory and research: Analyses from multiple paradigms* (pp. 231–240). Chicago: National Reading Conference.

Lindberg, C. A. (Ed.). (2002). *The Oxford American College Dictionary*. New York: G. P. Putnam's Sons.

Lindsey, K., Manis, F., & Bailey, C. (2003). Prediction of first-grade reading in Spanish-speaking English language learners. *Journal of Educational Psychology, 95*, 482-494.

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

Loftus, S. M. (2008). *Effects of a tier 2 vocabulary intervention on the word knowledge of kindergarten students at-risk for language and literacy difficulties* (Doctoral dissertation, University of Connecticut, 2008) ProQuest Dissertations and Theses. (UMI 3319089).

Lyman, H. B. (1978). *Test scores and what they mean (3rd ed.)*. Englewood Cliffs, NJ: Prentice Hall.

MacDonald, S., & McNaughton, S. (1999). Features of children's storytelling on entry to school. *New Zealand Journal of Educational Studies, 34*(2), 349-353.

Madsen, H. S., & Jones, R. L. (1981). Classification of oral proficiency tests. In A. S. Palmer, P. J. M Groot, & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 15- 30). Washington, DC: Teachers of English to Speakers of Other Languages (TESOL).

Manis, F., Lindsey, K., & Bailey, C. (2004). Development of reading in grades K-2 in Spanish-speaking English language learners. *Learning Disabilities Research & Practice, 19*, 214-224.

Marston, D. (2005). Tiers of intervention in responsiveness to intervention: Prevention outcomes and learning disabilities identification patterns. *Journal of Learning Disabilities, 38*(6), 539-541.

Mason, J. M., Stahl, S. A., Au, K. A., & Herman, P. A. (2003). Reading: Children's developing knowledge of words. In J. Flood, D. Lapp, J. R. Squire & J. M. Jensen (Eds.), *Handbook of research on teaching the English language arts* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

McMillan, J. H. (2007). *Classroom assessment: Principles and practice for effective standards-based instruction* (4th ed.). Boston, MA: Pearson Allyn and Bacon.

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in offshore assessments of speaking skills in occupational settings. *Language Testing, 14*(2), 140-156.

Mellard, D. F., Byrd, S. E., Johnson, E., Tollefson, J. M., & Boesche, L. (2004). Some effects of the nature and frequency of vocabulary instruction on the knowledge and use of words. *Reading Research Quarterly, 20*(4), 482-496.

Meltzoff, J. (1998). *Critical thinking about research*. Washington, DC: American Psychological Association.

Meyer, D. J. (2000). *Evaluating urban elementary bilingual classrooms through the four-dimensional transitional bilingual pedagogical theory.* Unpublished doctoral dissertation, Sam Houston State University, Huntsville, Texas.

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.

Mitchell, R., & Myles, F. (1998). *Second language learning theories*. London: Arnold Publishers.

Morrow, L. M. (1985). Retelling stories: A strategy for improving young children's comprehension, concept of story structure, and oral language complexity. *The Elementary School Journal 85*(5), 646-661.

Morrow, L. M. (1990). Assessing children's understanding of story through their construction and reconstruction of narrative. In L. M. Morrow & J. K. Smith (Eds.), *Assessment for instruction in early childhood* (pp. 110-134). Englewood Cliffs, NJ: Prentice Hall.

Munby, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.

Murdock, S. H. (2006). *Population change in Texas: Implications for human and socioeconomic resources in the 21st century*. Retrieved January 13, 2007, from http://txsdc.utsa.edu/presentations

Murphy, V. (1997). The effect of modality on a grammaticality judgement task. *Second Language Research, 13*, 34-65.

Muter, V., & Diethelm, K. (2001). The contribution of phonological skills and letter knowledge to early reading development in a multilingual population. *Language Learning, 51*(2), 187-219.

Naglieri, J. A. (1997). *Naglieri nonverbal ability test*. San Antonio, TX: The Psychological Corp.

Nagy, W. E., & Herman, P. (1985). Incidental vs. instructional approaches to increasing reading vocabulary. *Educational Perspectives, 23*, 16-21.

Nagy, W. E., & Herman, P. (1987). Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. In M. G. McKeown & M. E. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 19-35). Hillsdale, NJ: Lawrence Erlbaum.

Nambiar, M. K., & Goon, C. (1993). Assessment of oral skills: a comparison of scores obtained through audio recordings to those obtained through face-to-face evaluation. *RELC Journal, 24*(1), 15-31.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

National Center for Education Statistics. (2005). *Digest of education statistics: 2005*. Retrieved January 13, 2007, from http://nces.ed.gov/programs/digest/d05/tables/dt05_016.asp?referer=list

National Clearing House for English Language Acquisition and Language Instruction Educational Programs. (n.d.). *Woodcock Language Proficiency Battery - Revised* Retrieved July 11, 2007, from http://www.ncela.gwu.edu/pubs/eacwest/elptests.htm

National Institute of Child Health and Human Development. (2000). *Report of the national reading panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction (NIH Publication No. 00-4769)*. Washington, DC: U.S. Government Printing Office.

Nitko, A. J., & Brookhart, S. M. (2007). *Educational assessment of students* (5th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.

O'Neil, J. (1992). Putting performance assessment to the test. *Educational Leadership, 49*(8), 14-19.

Oosterhof, A. (2001). *Classroom applications of educational assessment*. (3rd ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Ovando, C. J., Collier, V., & Combs, M. C. (2003). *Bilingual and ESL classrooms: Teaching in multicultural contexts*. New York: McGraw-Hill Companies, Inc.

PacifiCorp Foundation. (2004). *Oregon project optimize*. Retrieved July 12, 2007, from http://www.pacificorpfoundation.org/Article/Article25116.html

Pappas, C. C., & Pettegrew, B. S. (1991). Learning to tell: Aspects of developing communicative competence in young children's story retellings. *Curriculum Inquiry, 21*(4), 419-434.

Paribakht, T. S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition*. New York: Cambridge University Press.

Paulston, C. B. (1990). Educational language policies of Utopia. In B. Harley, P. Allen, J. Cummins & M. Swain (Eds.), *The development of second language proficiency* (pp. 187-200). Cambridge: Cambridge University Press.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). New York: Holt, Rinehart, and Winston.

Pellegrini, A., & Galda, L. (1982). The effects of thematic-fantasy play training on the development of children's story comprehension. *American Educational Research Journal, 19*, 443-452.

Pickert, S. M., & Chase, M. (1978). Story retelling: An informal technique for evaluating children's language. *The Reading Teacher, 11*(4), 528-531.

Pienemann, M., & Johnson, M. (1987). Factors influencing the development of second language proficiency. In D. Nunan (Ed.), *Applying second language acquisition research* (pp. 45-144). Adelaide, Australia: National Curriculum Resource Centre.

Pinnell, G. S., & Jaggar, A. M. (2003). Oral language: Speaking and listening in elementary classrooms. In J. Flood, D. Lapp, J. R. Squire & J. M. Jensen (Eds.), *Handbook of research on teaching the English language arts* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Piper. (2003). *Language and learning: The home and school years* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Proctor, C., Carlo, M., August, D., & Snow, C. E. (2005). Native Spanish-speaking children reading in English: Toward a model of comprehension. *Journal of Educational Psychology, 97*, 246-256.

Quiros, A. M. (2008). *Structured story reading and retell related to listening comprehension and vocabulary acquisition among English language learners*. Unpublished doctoral dissertation, Texas A&M University, College Station, TX.

Ramírez, J. D., & R.T. International. (1992). Executive summary. *The Bilingual Research Journal, 16*(1-2), 1-62.

Raosoft. (n.d.). *Sample size calculator*. Retrieved March 9, 2007, from http://www.ezsurvey.com/samplesize.html

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1-32.

Read, J. A. S. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder, A. Brown, K. Grove, N. Hill, T. Iwashita, T. Lumley, T. McNamara & K. O. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (Vol. 11, pp. 82-96). Cambridge: Cambridge University Press.

Riches, C., & Genesee, F. (2006). Crosslinguistic and crossmodal issues. In F. Genesee, K. J. Lindholm, W. M. Saunders & D. Christian (Eds.), *Educating English language learners: A synthesis of research evidence* (pp. 64-108). New York: Cambridge University Press.

Riley, J., & Burrell, A. (2007). Assessing children's oral storytelling in their first year of school. *International Journal of Early Years Education, 15*(2), 181-196.

Roberts, G., Good, R., & Corcoran, S. (2005). Story retell: A fluency-based indicator of reading comprehension. *School Psychology Quarterly, 20*(3), 304-317.

Rosenfeld, S. A., & Shin, M. R. (1989). Mini-series on curriculum based assessment. *School Psychology Review, 18*, 299-370.

Rossell, C. H. (1992). Nothing matters?: A critique of the Ramírez, et al. longitudinal study of instructional programs for language-minority children. *The Bilingual Research Journal, 16*(1-2), 159-186.

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.

Saenz, L. M., Fuchs, L. S., & Fuchs, D. (2005). Peer-assisted learning strategies for English language learners. *Exceptional Children, 71*(3), 231-247.

Saunders, W. M., & O'Brien, G. (2006). Oral language. In F. Genesee, K. Lindholm-Leary, W. M. Saunders & D. Christian (Eds.), *Educating English language learners: A synthesis of research evidence*. New York: Cambridge University Press.

Savignon, S. J. (1983). *Communicative competence*. Reading, MA: Addison-Wesley.

Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising

predictors. In B. K. Shapiro, P. J. Accardo & C. A. J. (Eds.), *Specific reading disabilities: A review of the spectrum*. Timonium, MD: York Press.

Schrank, F. A., Fletcher, T. V., & Alvarado, C. G. (1996). Comparative validity of three English oral language proficiency tests *The Bilingual Research Journal 20*(1), 55-68.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17*, 229-255.

Shapiro, E. S. (1990). An integrated model for curriculum based assessment. *School Psychology Review, 19*, 331-349.

Shepard, L. A. (1995). Using assessment to improve learning. *Educational Leadership, 52*(5), 38-43.

Shepaz, L. (1991). Will national tests improve student learning? *Phi Delta Kappan, 73,* 232-238.

Sheskin, D. J. (2007). *Handbook of parametric and nonparametric statistical procedures* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.

Shohamy, E. (2001). *The power of tests: A critical perspective on the use of language tests*. London: Pearson Education.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effects of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33.

Sinclair, J. M. (1966). Beginning the study of lexis. In C. E. Bazell, J. C. Catford, M. A. K. Halliday & R. H. Robins (Eds.), *In memory of J. R. Firth* (pp. 410-430). London: Longman.

Singleton, D. (1999). *Exploring the second language mental lexicon*. Cambridge: Cambridge University Press.

Skehan, P. (1989). *Individual differences in second-language learning*. London: Edward Arnold.

Skehan, P. (1991). Progress in language testing: the 1990s. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 3–21). London: Macmillan Publishers Limited.

Skinner, B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Slavin, R. E., & Madden, N. A. (Eds.). (2006). *One million children: Success for All*. Thousand Oaks, CA: Corwin Press.

Snow, C., Cancino, H., de Temple, J., & Schley, S. (1991). Giving formal definitions: A linguistic or metalinguistic skill. In E. Bialystok (Ed.), *Language processing in bilingual children* (pp. 90-112). Cambridge: Cambridge University Press.

Snow, C. E. (1983). Literacy and language relationships during the preschool years. *Harvard Educational Review, 53*(2), 165-189.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* National Research Council. Washington, DC: National Academy Press.

Snow, C. E., Cancino, H., Gonzalez, P., & Shriberg, E. (1987). *Second language learners' formal definitions: An oral language correlate of school literacy (Report No. CLEAR-TR5)*. Washington, DC: Office of Educational Research and Improvement.

Stahl, S. A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*(1), 72-110.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*(4), 360-406.

Stewig, J., & Young, L. (1978). An exploration of the relation between creative drama and language growth. *Children's Theatre Review, 27*, 10-12.

Stiggins, R. (1999). Assessment, student confidence, and school success. *Phi Delta Kappan,81*(3), 191-198.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. A. Gass & C. G. Madden (Eds.), *Input in second language acquisition* (pp. 235-253). Rowley, MA: Newbury House.Bilingual Education and Special Language Programs, Subchapter B 29.053 (1995).

Texas Education Agency. (2004). *Academic excellence indicator system: 2003-2004 state performance report.* Retrieved January 28, 2006, from http:www:tea.state.txus/perfreport/aeis/2004/state.html

Texas Education Agency. (2006a). *Academic excellence indicator system: 2005-2006 state performance report*. Retrieved January 13, 2007, from http://www.tea.state.tx.us/perfreport/aeis/2006/state.html

Texas Education Agency. (2006b). *Aldine ISD named TAPE recipient*. Retrieved July 4, 2007, 2007, from http://www.tea.state.tx.us/comm/stars/links_pdfs/0506/aldineISD051906.pdf#xml=http://www.tea.state.tx.us/cgi/texis/webinator/search/xml.txt?query=aldine&db=db&id=30c86a727415fc5b

Thomas, W. P. (1992). An analysis of the research methodology of the Ramírez study. *The Bilingual Research Journal, 16*(1-2), 213-245.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach*. New York: The Guilford Press.

Thompson, B. (Ed.). (2003). *Score reliability: Contemporary thinking on reliability issues*. Newbury Park, CA: Sage.

Tinajero, J. V., & Hurley, S. R. (2001). Assessing progress in second-language acquisition. In S. R. Hurley & J. V. Tinajero (Eds.), *Literacy assessment of second language learners* (pp. 31-40). Needham Heights, MA: Allyn & Bacon.

Tinajero, J. V., & Schifini, A. (1997). *Into English! Level B, teacher's guide.* Carmel, CA: Hampton-Brown Books.

Tong, F. (2006). *Oral English development and its impact on emergent reading achievement: A comparative study of transitional bilingual and structured English immersion.* ProQuest Digital Dissertation database (UMI No. 3296556.

Tong, F., Irby, B., Lara-Alecio, R., & Mathes, P. G. (2008a). English and Spanish acquisition by Hispanic second-graders in developmental bilingual programs: A 3-year longitudinal randomized study. *Hispanic Journal of Behavioral Sciences, 30*(4), 500-529.

Tong, F., Lara-Alecio, R., Irby, B., Mathes, P. G., & Kwok, O. (2008b). Accelerating early academic oral English development in transitional

bilingual and structured English immersion programs. *American Educational Research Journal, 45*(4), 1011-1044.

U.S. Department of Education. (2003). *No Child Left Behind*. Retrieved July 23, 2007, from http://www.nochildleftbehind.gov/next/overview/index.html

U.S. Department of Education. (2004). *Language minorities and their educational and labor market indicators -recent trends: Statistical analysis report* (No. NCES 2004-009): National Center for Education Statistics.

University of Oregon Center on Teaching and Learning. (n.d.). *DIBELS Measure Download Page - 6th edition*. Retrieved March 31, 2007, from https://dibels.uoregon.edu/measures/download.php

Vaughn, S., Mathes, P. G., Linan-Thompson, S., Cirino, P. T., Carlson, C. D., Pollard-Durodola, S. D., et al. (2006). Effectiveness of an English intervention for first-grade English language learners at risk for reading problems. *The Elementary School Journal, 107*(2), 153-180.

Verhallen, M., & Schoonen, R. (1993). Vocabulary knowledge of monolingual and bilingual children. *Applied Linguistics, 14*, 344-363.

Verhoeven, L., & Vermeer, A. (1992). Modeling communicative second language competence. In L. Verhoeven & E. de Jong (Eds.), *The construct of language proficiency*. Philadelphia: John Benjamin's Publishing Company.

Walichowski, M., Irby, B., & Pollard-Durodola, S. (2007). [Semantic and Syntactic Scoring System (S4)]. Unpublished curriculum-based assessment. College Station, TX: Texas A&M University, Bilingual Programs.

Walsh, W. B., & Betz, N. E. (2001). *Tests and Assessment (4th ed.)*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Watson, L. T., Layton, P. P., & Abraham, L. (1994). Enhancing emerging literacy in a language preschool. *Language, Speech, and Hearing Services in Schools, 25*, 136-145.

Whaley, J. (1981). Story grammars and reading instruction. *Reading Teacher, 34*, 762-771.

Wiig, E. H., Secord, W., & Sernel, E. (1992). *Clinical evaluation of language fundamentals - preschool UK (CELF - preschool UK)*. London: London Psychological Corp.

Wolfinger, R., Tobias, R., & Sall, J. (1994). Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal on Scientific Computing, 15*(6), 1294-1310.

Woodcock, R. W. (1991a). *Woodcock language proficiency battery-revised*. Circle Pines, MN: American Guidance Service.

Woodcock, R. W. (1991b). *Woodcock language proficiency battery-revised: Examiner's manual*. Itasca, IL: Riverside Publishing.

Zimiles, H., & Kuhns, M. (1976). *A developmental study of the retention of narrative material, final report*. New York: Bank Street College of Education. (ERIC Document Reproduction Service No. ED 160978).

Zimmerman, C. B. (1997). Historical trends in second language vocabulary instruction. In J. Coady & T. Huckin (Eds.), *Second language vocabulary instruction* (pp. 5-19). New York: Cambridge University Press.

APPENDIX A

PROJECT STELLA VOCABULARY FLUENCY MEASURE (Use Word in a Sentence)
Protocol: With L1 Clarifications/Modifications

School		**_____C I**
Face		**_____C I**
Hop		**_____C I**
Climb		**_____C I**
Mittens		**_____C I**
Caterpillar		**_____C I**
Born		**_____C I**
Feathers		**_____C I**
Woods		**_____C I**
Scarf		**_____C I**
Munch		**_____C I**
Swooped		**_____C I**
Spring		**_____C I**
Crowd		**_____C I**
Squirm		**_____C I**
Shelter		**_____C I**
Perch		**_____C I**
Trail		**_____C I**

APPENDIX B

SEMANTIC + SYNTACTIC SCORING SYSTEM (S4)

1 NO KNOWLDEGE
No Knowledge of word meaning, incorrect response, code-switching, response in Spanish or in any other language that is not English, student merely repeats the target word, or over use of a stem (example: I see *cat*. I see *dog*. I see *library*, I see *book*). There is some indication that the student ***many not know*** the word meaning. One could infer that the student does not know the meaning of the word based on the response provided.

2 SOME KNOWLEDGE
Partial or incomplete knowledge of word meaning with or without syntactic errors (examples: Cars are *traffic*. Face *freckles*.). Also, appropriate word associations (example: the target word is *milk* and the student responds cow). Students demonstrate some knowledge of the target word but do not possess enough knowledge of English syntax to respond with language that is more elaborate.

3 KNOWLEDGE + SIMPLE SENTENCE (SUBJ + VERB OR SUBJ. +VERB+OBJECT)
Target word used in an appropriate and meaningful context (example: I have two *feets* or *foots*. I wear *boots,* I can *stand*). There is a complete thought. Syntactic errors do not interfere with conveying a complete thought. Sentence elaboration is limited to the use of determiners such as: the, a, an, etc.)

4 KNOWLEDGE + ELABORATE SENTENCES
Target word used in an appropriate meaningful context with an elaborate syntactic structure (example: I like to play at the *beach* in the sand). There is use of advanced and sophisticated language. Student might extend context beyond self (example: My mother has a *purse.* My teacher has a big *desk*. The baby can eat baby *food*.). Syntax supersedes SUBJ +VERB+OBJ. Elaboration is determined by the use of the following syntactic structures and goes beyond the use of determiners:

- Prepositional phrases (at the beach, on the table)
- Compound objects (tall and slim; cake and ice cream)
- Modifiers (green grass, fuzzy hair, cold wind)

Consider
Primary focus is on KNOWLEDGE of target words followed by the ability to use appropriate SYNTAX. Syntax may or may not impede the ability to express knowledge. Think, “Is the item closer to being rated as a 1 or 2, a 2 or 3, a 3 or 4. When in doubt:

a. examine the student’s knowledge if the word (complete vs. incomplete thought)
b. examine the syntax of the sentence (simple subj/v/o vs. use of modifiers etc.)

APPENDIX C

Practice A: Distinguishing between a 2 and a 3

_____ 1. the catch *ball*

_____2. I catch *ball*

_____ 3. *flower* yellow

_____ 4. the *flower* is yellow

_____ 5. I *smart*

_____6. I am *smart*

_____7. a *lunch* for a eat

_____8. eat you *lunch*

_____9. happy *face*

_____10. my *face* happy

Examples of Category 2:

1. **the green *grass***
2. **those skinny *legs***
3. **on the big *bus***
4. **the *wood* door**
5. **yellow *bird* is**

Examples of Category 3:

1. **I see green *grass***
2. **those are skinny *legs***
3. **he ride the *bus***
4. **the door is *wood***
5. **the *bird* is yellow**

APPENDIX D

Practice A: Distinguishing between a 3 and a 4

_____ 1. *school is* fun

_____2. I like to go to *school*

_____ 3. the *park has a swing*

_____ 4. the park has a big *swing*

_____ 5. I like *swim*

_____6. he *swim* fastest

_____7. I am *scared* at night

_____8. I am *scared*

_____9. my *baby* brother eat baby food

_____10. I have a *baby* brother

Examples of Category 3:

1. **I am *nice***
2. **the *bird* has wings**
3. ***trucks* is big**
4. **the door is *wood***
5. **the *bird* is yellow**

Examples of Category 4:

1. **I sweet and *nice* girl**
2. **the *bird* lives up in tree**
3. **He drived a big truck**
4. **The door is made with *wood***
5. **The *bird* flying up in sky**

APPENDIX E

_____ 1. **I see *drink***

_____ 2. **I see *bus*, I see *chair*, I see *books***

_____ 3. **I make *basket***

_____ 4. **I like to *slide* down**

_____ 5. **I buy it at the *store***

_____ 6. **fruit *apple***

_____7. **my *heart* beats fast when I run**

_____ 8. ***chair* to sit**

_____ 9. **my brother is a *trip* for he make jokes**

_____ 10. **I like *clouds***

_____ 11. **I like white *clouds***

_____ 12. **I like eat the blueberry *plates***

_____ 13. **you *eat***

_____ 14. **the *horse* wears a hat**

_____ 15. **a baby sleep *crib***

_____ 16. **the *bee* buzes around**

_____ 17. **the *clown* funny**

_____ 18. **the baby sits up**

_____ 19. **night (the word given to the student was *sleep*).**

APPENDIX F

_____ 1. **library *books***

_____ 2. **I have *two* feets**

_____ 3. **jump is to *hop***

_____ 4. **I like to *slide* down**

_____ 5. **I like *boots*. I like *cats*. I like *dogs*.**

_____ 6. **the *bug* is small and eats grass**

_____7. **my mommy's *coat* is fur**

_____ 8. **the *coat* is fur**

_____ 9. **I eat *trees***

_____ 10. **my *hat* is big and yellow**

_____ 11. **I see *invisible***

_____ 12. **a like to *dive* in the pool**

_____ 13. **I like *fall* and *winter***

_____ 14. **my mother does not like *traffic***

_____ 15. **a baby can *crawl***

_____ 16. **big *house***

_____ 17. **I *stand* in the line in the cafeteria**

_____ 18. **es un camino para los carros**

_____ 19. **I like to go *camping***

APPENDIX G

SEMANTIC + SYNTACTIC SCORING SYSTEM

0 NO RESPONSE

No response was given, at all. The response was entirely in Spanish.

1 NO KNOWLDEGE

There is some indication that the student does not or may not know the word meaning. Based on the response, one may infer that the student doesn't know meaning of the word.

- Any code-switching
- Incorrect Response
- Target word was merely repeated (EXAMPLE: the target word is *house* and the student says *house*)
- Repetitive, over use or consecutive use of a stem (EXAMPLE: I see *cat*. I see *dog*. I see *library*. I see *book*.)

2 SOME KNOWLEDGE

Partial or incomplete knowledge of word meaning with or without syntactic error. Students demonstrate some knowledge of the target word but do not possess enough knowledge of English syntax to respond with more elaborate language. If the student does not demonstrate correct knowledge of the word then they do not fall in this category, they would be considered a 1.

- Student makes a correct, single-word association (EXAMPLE: the target word is *milk* and the student just responds *cow*)
- Student uses more than one word, but it is still just a correct association (EXAMPLE: Cars are traffic. Face freckle.)

3 KNOWLEDGE + SIMPLE SENTENCE (SUBJ + VERB OR SUBJ. +VERB+OBJECT)

There may be syntactic errors, but they do not hinder the student from conveying a complete thought.

- There may be a use of simple determiners (the, a, an, etc.) (EXAMPLE: The boy *runs*. I have a *cat*.) Or the determiner might be missing, but the thought is still clear.
- Syntactic errors (if present) do not interfere with the conveying of word knowledge and thought. (EXAMPLE: The boy runned. I have two *feets*).
- Target word is used in an appropriate and meaningful context. (EXAMPLE: The cow makes the *milk.)*.
- There is a complete thought (EXAMPLE: I can *stand*.)

4 KNOWLEDGE + ELABORATE SENTENCES

Target word used in an appropriate meaningful context with an elaborate syntactic structure. Use of more advanced and sophisticated language. Syntax supersedes SUBJ –VERB-OBJ. (EXAMPLE: I like to play at the beach because I like sand.)

- Elaboration goes beyond the use of determiners and should include one or more of the following:
  - Prepositional phrases (at the beach, on the table)
  - Compound objects (tall and slim; cake and ice cream)
  - Modifiers (green grass, fuzzy hair, cold wind)
  - Modifiers beyond self (my mother, my teacher, his brother, her cat, and etc) reference to someone that is not the student, the student goes beyond "I, me, my," in addition to one of the above components.

Note:

- Primary focus is on KNOWLEDGE of target words followed by the ability to use appropriate SYNTAX. Syntax may or may not impede the ability to express knowledge. Think, "Is the item closer to being rated as a 1 or 2, a 2 or 3, a 3 or 4. When in doubt:

a. examine the student's knowledge of the word (complete vs. incomplete thought)
b. examine the syntax of the sentence (simple (sub/v/o) vs. use of modifiers, etc.

- Each response should be considered independent from the others (except when a student us using repetitive and consecutive stems).
- If children repeat a word as part of processing do not assume that is incorrect word knowledge of incorrect syntax (e.g. "the boy, the boy, the boy ran.) In this spoken text we do not count against hesitations, unfilled pauses (nothing is said during a pause), filled pauses (uh, um, mm, etc.), repetitions, or false starts.

APPENDIX H

Progression Chart	School	Face	Hop	Climb	Mittens	Caterpillar	Born	Feathers	Woods	Scarf	Munch	Swooped	Spring	Crowd	Squirm	Shelter	Perch	Trail
No Response (if you checked one or more do not proceed and assign a 0 for that word)																		
No Answer Given or None of it in English																		
If any checked - Stop & assign (0) pts.																		
No Knowledge (if you checked one or more do not proceed and assign a 1 for that word)																		
Code-Switching																		
Incorrect Reponse																		
Repeated Target Word (and can't be rated)																		
Repetitive Stem Use																		
If any checked - Stop & assign (1) pt.																		
Some Knowledge (if you checked two or more, please see if you can proceed to the next category)																		
Shows Partial or Incomplete BUT Correct Knowledge																		
or Shows Complete & Correct Knowledge																		
Word Association (single or phrase)																		
Syntax errors BUT they do not hinder response																		
or No Syntax Errors																		
See if you can progress - if no, assign (2) pts.																		
K + Simple Sentence (if you checked two or more, please see if you can proceed to the next category)																		
Is there a subject & verb																		
Is there a subject & verb & object																		
Syntax errors BUTthey do not hinder response																		
or No Syntax Errors																		
Complete Thought																		
Context is appropriate																		

See if you can progress - if no, assign (3) pts.

K+ Elaborate Sentences ***(if you checked any here, then the score is a 4)***

may include prepositional phrases

or may include compound (subj., pred., or object)
or may include modifiers (adv & adj)
or may have many details
if any of these, assign (4) pts.

Total Score

APPENDIX I

Target Word	Student Transcribed Response	Score 0, 1, 2, 3, or 4	Score? - possibilities?	Justify your final decision. Use the scoring system to articulate and finalize your decision.
School	I go to school			
Face	my face is pretty			
Hop	what is hop?			
Climb	I go to the climb			
Mittens	Mittens			
Caterpillar	I eat			
Born	I'm born			
Feathers	n/r			
Woods	woods, woods they sport			
Scarf	the boy gots the scarf			
Munch	n/r			
Swooped	swoop, the girl is eating soup			
Spring	spring, what is spring			

Crowd	Uhm			
Squirm	n/r			
Shelter	I in shelter			
Perch	my mom got a purse			
Trail	n/r			
	Total Score			

APPENDIX J

Target Word	Student Transcribed Response	Score 0, 1, 2, 3, or 4	Score? – possibilities?	Justify your final decision. Use the scoring system to articulate and finalize your decision.
School	the school is for we speak English			
Face	N/R			
Hop	the hop is for reduce it			
Climb	we climb in the stairs			
Mittens	the mittens we use them we speak English			
Caterpillar	the caterpillar is green			
Born	I don't like anything born			
Feathers	the feathers are for we use cause we like them			
Woods	the woods the woods we like them cause they are beautiful			
Scarf	the scarf uhm we don't need it because we scratch ourself			
Munch	we munch in the story			
Swooped	we swooped in the park			
Spring	we like the sprung cause it is so beautiful			

Crowd	we crowd ourself			
Squirm	We squirm in our hand			
Shelter	we shelter in my room			
Perch	we perch in the library			
Trail	we trail in the classroom			
	Total Score			

APPENDIX K

Target Word	Student Transcribed Response	Score 0, 1, 2, 3, or 4	Score ? - possibilities?	Justify your final decision. Use the scoring system to articulate and finalize your decision.
School	Los niños van a la escuela			
Face	Eyes			
Hop	Los niños brincan en la cama			
Climb	I see a climb			
Mittens	I see a mittens			
Caterpillar	I see a caterpillar			
Born	I see a born			
Feathers	feathers is that birds fly			
Woods	woods are uhm Indian so they can make boats			
Scarf	Scarf			
Munch	Lunch			
Swooped	n/r			
Spring	in the spring, in the spring, in the spring are lots of flowers			

Crowd	crowd is when there is something and you get everything together			
Squirm	the squirm lives in the tree			
Shelter	Home			
Perch	in the perch is a place who can birds can climb on			
Trail	trail is when the horse goes by the trail			
	Total Score			

APPENDIX L

Oral Language Proficiency and Expressive Vocabulary

The link between oral language proficiency and vocabulary knowledge to literacy and academic success for children, particularly for second languages learners is incontrovertible. It is important that teachers be able to ascertain the oral proficiency level and vocabulary knowledge of each student. These data can be used to inform instruction and provide a basis for differentiated instruction.

Rationale for STELLA Vocabulary Fluency Measure

The Semantic and Syntactic Scoring System (S4) is the scoring instrument for the STELLA Vocabulary Fluency Measure. The STELLA Vocabulary Fluency Measure was an assessment created to test oral proficiency and expressive vocabulary knowledge. The instrument is composed of 18 target vocabulary words. These words have been taught in the classroom directly and indirectly. In this particular case, the words were taught in kindergarten, Bilingual and ESL classrooms. This test was administered individually by paraprofessionals. The paraprofessional pulled children from class, one-at-a-time, and took them to a quiet room for the test. The student was instructed to provide an English sentence to each word that they were given. These responses were recorded with a tape recorder. Then the taped responses where transcribed for rating with the S4.

Rationale for Syntactic and Semantic Scoring System (S4)

Because a student's ability to demonstrate expressive vocabulary knowledge is limited by his or her oral proficiency and his or her ability to demonstrate oral proficiency is limited by his or her word knowledge, it becomes important to use an instrument that accounts for both of these areas and attempts a more holistic/integrated approach to assessment. Furthermore, commercialized instruments should not be the sole means for testing oral proficiency or vocabulary because these tests offer a panoramic assessment; they are not focused on the curriculum that is being taught in the class. Rarely, do scores on commercialized tests of oral proficiency or vocabulary have a direct connection and influence on the curriculum. Therefore, an assessment that utilizes vocabulary words from the curriculum can offer insight and deeper understanding of a student's performance and progress within the context of the classroom. With this instrument, teachers will be able to use their own target words to assess word knowledge and oral proficiency of each individual student.

Initial (First-time) Rater Training

To ensure accuracy and efficiency in using the S4 raters should do the following:

1. Read this manual.
2. Read the Semantic and Syntactic Scoring System explanation of categories (Appendix 1) and refer to it as they are scoring sentences.
3. View the Progression Chart (Appendix 2).
4. Rate the sample sentences provided in Word and Sentence Table A (Appendix 3) using the Progression Table (A) (Appendix 4).
5. Compare responses, from step #4, with the scores and explanation of rating in Appendix 5.
6. If there is any discrepancy between the scores provided by the rater and the scores provided in this manual for step #4 then the rater should review the above materials in order to understand why the discrepancy occurred.
7. If the scores concurred, at or above 95%, then the rater is prepared to proceed with Word and Sentence Table B (Appendix 6) and Progression Table (Appendix 7). Again, consistency in scoring should be at or above 95%. If they are then the rater is prepared to use the instrument.

Scoring Considerations

Each sentence is treated as a separate sentence and should not be scored in comparison to others in the table or to responses given in other tests by other students. The *ONLY* exception to this is when the student has used a repetitive stem, within the same given test, the sentences prior and after the sentence in question will need to be evaluated to see if the student is using a repetitive stem. A repetitive stem cannot be determined in isolation. An example of a repetitive stem is "*The girl likes **cars*** .*" "*The girl likes **books.**"* *The girl likes **run**."* In this case, each sentence will receive a 1 because we cannot be sure that they student really knows what the words mean. If the student had said, "the girl likes cars because she wants to drive them" and "the girl likes books because she wants to read" and "the girl likes to run because she likes to exercise" then we would not consider this a stem because each sentence is a stand-alone sentence and shows that the student knows how to use the word appropriately.

* The bold word represents the target word.

Scoring Procedures
The rater will use the sentences provided in the Word and Sentence Tables (or a similar table if adapting it for the classroom). The Word and Sentence Table has the transcribed sentences that kindergarten students produced. Also, a blank progression table is needed for each test.

Determine Score
The Progression Table is used to score each student's test. The rater starts with the first word and reads the respective sentence (from the Word Sentence Table). Then the rater starts at the top with **Category O (No Response)** and checks off any criteria that the student meets. Then the rater will keep moving down to the next category. If the rater gets to a category and realizes that the response does not meet any criteria for that category, then the rater will go back to the above category and that will be the best score for the response. Each word (response) will be a 0, 1, 2, 3, or 4. The raw score for the students is the aggregate of all the scores given for each word. The progression chart, if used correctly, will guide the rater. Some of the differences that the raters needs to be aware of are printed in the progression chart and restated in **Category Notes and Definitions**.

Category Notes and Definitions:

Category ***No Response (0 Points):*** if anything is checked in this category then stop, the sentence can only rate as a 0. In this category the student did not say anything, at all. The student could have responded ***entirely*** in a language other than English. Or the student made a comment like "what is that" "I don't know." (If the student says, "**what is *(insert target word)*?"** Then they are considered to have repeated the word and that belongs in the next category).

Category ***No Knowledge(1 Point):*** if one or more is checked here then stop, the sentence can only be rated as a 1.

- Code-switching is when some words were in English and other words were in another language within the given sentence.
- Incorrect Response means that the response is not correct, not plausible. For example, if the student says, "I can eat a ***hop***" clearly the word is not used correctly. However, if a student says "my horse wears a ***hat***, "although horses do not traditionally wear hats, they could wear hats in a fictional story or in one's

imagination. *When teachers are the raters, it facilitates this aspect of scoring because they know the context in which the words were taught directly and indirectly and the scope of possible answers.*

- Repeated Target Word means that student merely repeated the word or used the word to inquire about it or state that they do not know it. If the word is ***swim***, the student might have said, "***swim***," "what is ***swim***," "I do not know ***swim***," "uhm, uhm, uh ***swim, swim*** is, uh…"
- Repetitive Stem Use is when the 3 or more consecutive sentences use the same sentence stem and the only difference among the sentences is the target word. For example, if the target words are ***bike, snow, dance*** and the student's sentences are similar to "I like ***bike,***" I like ***snow,***" and I like ***dance,***" then these are rated as a 1 – we give them credit for repeating the word but nothing more.

Category ***Some Knowledge(2 Points):*** In this category you will always check either "*Shows Partial or Incomplete BUT Correct Knowledge"* or *"Shows Complete and Correct Knowledge"* **AND** *"Syntax errors BUT they do not hinder response"* or *"No Syntax Errors"* because it will always be one or the other for each. The key here is to see whether the response is merely and association or a sentence.

- Word Association (phrase or word) means that the student did not provide a complete sentence but they did state something that shows that they know an association for the word. If the word is ***snow*** and they respond ***cold*** or ***it's cold*** that is an association. If the student had said, *"the* ***snow*** *is cold"* or even *"****snow*** *is cold"* then these statements are beyond a mere associations and they are considered sentences and should receive a higher rating.

Category ***Knowledge + Simple Sentence(3 Points):*** In this category you will always check either *"Is there a subject & verb" or "Is there a subject & verb & object"* **AND** *"Syntax errors BUT they do not hinder response"* or *"No Syntax Errors"* because it will always be one or the other for each. The key here is to see whether the response is merely and association or a sentence.

- Complete thought and Context Appropriate means that the responses is a well conveyed sentence it may or may not have errors, but the errors are minimal or there are minimal omissions that do not hinder you from understanding the intent of the response.

Category ***Knowledge + Elaborate (4 Points):*** In this category you subsume that the above category (Knowledge + Simple Sentence) was met. Here we are testing to see if we can go beyond that category and into category 5. If a sentence has not made it through category 4 it can't be considered for category 5.

- Elaborate Sentences means that the sentence includes any or some of the following prepositional words/phrases, compounds (subject, predicate, or object), modifiers (adjectives and adverbs), and details. Example would be, with the target word jump, *“I like to **jump**”* is a category 4 and what would make it a category 5 could be, *“I like to **jump** over the box.”*

SEMANTIC + SYNTACTIC SCORING SYSTEM

0 NO RESPONSE

No response was given, at all. The response was entirely in Spanish.

1 NO KNOWLDEGE

There is some indication that the student does not or may not know the word meaning. Based on the response, one may infer that the student doesn't know meaning of the word.

- Any code-switching
- Incorrect Response
- Target word was merely repeated (EXAMPLE: the target word is *house* and the student says *house*)
- Repetitive, over use or consecutive use of a stem (EXAMPLE: I see *cat*. I see *dog*. I see *library*. I see *book*.)

2 SOME KNOWLEDGE

Partial or incomplete knowledge of word meaning with or without syntactic error. Students demonstrate some knowledge of the target word but do not possess enough knowledge of English syntax to respond with more elaborate language. If the student does not demonstrate correct knowledge of the word then they do not fall in this category, they would be considered a 1.

- Student makes a correct, single-word association (EXAMPLE: the target word is *milk* and the student just responds *cow*)
- Student uses more than one word, but it is still just a correct association (EXAMPLE: Cars are traffic. Face freckle.)

3 KNOWLEDGE + SIMPLE SENTENCE (SUBJ + VERB OR SUBJ. +VERB+OBJECT)

There may be syntactic errors, but they do not hinder the student from conveying a complete thought.

- There may be a use of simple determiners (the, a, an, etc.) (EXAMPLE: The boy *runs*. I have a *cat*.) Or the determiner might be missing, but the thought is still clear.
- Syntactic errors (if present) do not interfere with the conveying of word knowledge and thought. (EXAMPLE: The boy runned. I have two *feets*).
- Target word is used in an appropriate and meaningful context. (EXAMPLE: The cow makes the *milk.)*.
- There is a complete thought (EXAMPLE: I can *stand*.)

4 KNOWLEDGE + ELABORATE SENTENCES

Target word used in an appropriate meaningful context with an elaborate syntactic structure. Use of more advanced and sophisticated language. Syntax supersedes SUBJ –VERB-OBJ. (EXAMPLE: I like to play at the beach because I like sand.)

- Elaboration goes beyond the use of determiners and should include one or more of the following:
  - Prepositional phrases (at the beach, on the table)
  - Compound objects (tall and slim; cake and ice cream)
  - Modifiers (green grass, fuzzy hair, cold wind)
  - Modifiers beyond self (my mother, my teacher, his brother, her cat, and etc) reference to someone that is not the student, the student goes beyond "I, me, my," in addition to one of the above components.

Note:

- Primary focus is on KNOWLEDGE of target words followed by the ability to use appropriate SYNTAX. Syntax may or may not impede the ability to express knowledge. Think, "Is the item closer to being rated as a 1 or 2, a 2 or 3, a 3 or 4. When in doubt:

c. examine the student's knowledge of the word (complete vs. incomplete thought)
d. examine the syntax of the sentence (simple (sub/v/o) vs. use of modifiers, etc.

- Each response should be considered independent from the others (except when a student us using repetitive and consecutive stems).
- If children repeat a word as part of processing do not assume that is incorrect word knowledge of incorrect syntax (e.g. "the boy, the boy, the boy ran.) In this spoken text we do not count against hesitations, unfilled pauses (nothing is said during a pause), filled pauses (uh, um, mm, etc.), repetitions, or false starts.

S4 Practice Table (A)

These sentences are sample sentences selected from kindergarten children in Bilingual and Structured English Immersion classes. Normally, this table would reflect the responses given by just one child, but because this is for training purposes, it is important that the sentences selected reflect possible response variations.

Word	Sentence	Score
School	boys and girls is in the school	
Face	my face is white	
Hop	hop	
Climb	I climb	
Mittens	a boy use a mittens	
Caterpillar	I see a caterpillar	
Born	nacer	
Feathers	un bird tiene las feathers	
Woods	yes the trees are	
Scarf	neck	
Munch	a uhm carrots is for the munching of bunny rabbit and the the	
Swooped	no response	
Spring	the the the flowers and spring	
Crowd	a crowd is a big hat the queen wears on her head	
Squirm	a squirm was sitting	
Shelter	a boy was shelter	
Perch	a boy was perch	
Trail	a boy was trail	

Progression Chart	School	Face	Hop	Climb	Mittens	Caterpillar	Born	Feathers	Woods	Scarf	Munch	Swooped	Spring	Crowd	Squirm	Shelter	Perch	Trail
No Response *(if you checked one or more do not proceed and assign a 0 for that word)*																		
No Answer Given or Not in English																		
If any checked - Stop & assign (0) pts.																		
No Knowledge *(if you checked one or more do not proceed and assign a 1 for that word)*																		
Code-Switching																		
Incorrect Response																		
Repeated Target Word (and can't be rated)																		
Repetitive Stem Use																		
If any checked - Stop & assign (1) pt.																		
Some Knowledge *(if you checked two or more, please see if you can proceed to the next category)*																		
Shows Partial or Incomplete BUT Correct Knowledge																		
or Shows Complete & Correct Knowledge																		
Word Association (single or phrase)																		
Syntax errors BUT they do not hinder response																		
or No Syntax Errors																		
See if you can progress - if no, assign (2) pts.																		
K + Simple Sentence *(if you checked two or more, please see if you can proceed to the next category)*																		
Is there a subject & verb																		
Is there a subject & verb & object																		
Syntax errors BUT they do not hinder response																		
or No Syntax Errors																		
Complete Thought																		
Context is appropriate																		
See if you can progress - if no, assign (3) pts.																		

K+ Elaborate Sentences *(if you checked any here, then the score is a 4)*

may include prepositional phrases
or may include compound (subj., pred., or object)
or may include modifiers (adv & adj)
or may have many details

if any of these, assign (4) pts.

Answer and Explanations for S4 Practice Table (A)

Word	Sentence	Score	Explanation
School	boys and girls is in the school	4	Correct Use of Word Syntax error BUT I can understand clearly the message Compound subject (boys and girls) Preposition (in)
Face	my face is white	3	Clearly a 3 – tried to move on to 4 but it did not meet those requirements so went back
Hop	hop	1	All the student did was repeat the target word
Climb	I climb	3	May seem like a 2, but as I went on, it met most of 3 requirements – it is correct and it is a complete thought
Mittens	a boy use a mittens	3	Meets all the requirements of 3 – mistake with syntax but the sentence is understood
Caterpillar	I see a caterpillar	3	I can't know for sure if the child knows what a caterpillar is with this sentence BUT it is not a repetitive stem, it is correct, and we give the benefit of the doubt. If the child said "I see a caterpillar becoming a butterfly" then I know that they know "caterpillar" BUT again there is nothing wrong with the sentence and we can't PROVE that they do not know.
Born	nacer		In Spanish and we cannot go on, has to be a 0
Feathers	un bird tiene las feathers	1	Code-switch to another language other than English can't go on – has to be 1. Spanish reponses (or other languages) do not count BUT in this case some of the sentence was in English, too so the sentence is above a 0.
Woods	yes the trees are	2	I could not go on – it is an association woods and trees go together – although there are more words they do nothing – in essence all we can gather is that the student knows that trees and woods go together But we can't call this a sentence or a

			complete thought – we can only give credit for having an association
Scarf	neck	2	Again, a scarf goes with neck and that is all I can give credit for…it is correct but it is not a sentence
Munch	a uhm carrots is for the munching of bunny rabbit and the the	4	Here I could not decide it was partial or complete knowledge, so I checked off both – it does not affect the score, but it helps my thinking. Do not be confused by the excess words "uhm" "the, the" these are clearly words that the child uttered in trying to process his/her thoughts. We do not count off for that – try ignoring some of them and see if you can better judge the sentence.
Swooped	no response	0	The child did not say anything
Spring	the the the flowers and spring	2	Like in munch – ignore the "the, the, the" then you will see that this is just an association, it is not a sentence BUT it is a correct association "flowers and spring" do go together. If the child had said "eat and spring" then I could say that the child does not know what spring means. But with flower and spring – I can not really say that.
Crowd	a crowd is a big hat the queen wears on her head	1	As I go down the column, all I can do is check off incorrect response and once you check that, you can't go on. Although, this is a good sentence, the child thought the word was "crown" and it was "crowd" he does not get any credit beyond 1 point.
Squirm	a squirm was sitting	1	Again, like above I need to stop at 1 – a "squirm" can't sit …the response is incorrect. I can only give 1 point.
Shelter	a boy was shelter	1*	It is incorrect for starters, then I glance at the sentences that follow and see that the student is using a repetitive stem – these are all "a boy was" so the only category it meets is 1.
Perch	a boy was perch	1*	Same as above. If this sentence were "a boy was sitting on a perch" – then it would be okay, even with the "a boy was" because I can see that the student knows perch and they made

			the sentence different.
Trail	a boy was trail	1*	Same as with “shelter.” If the student said “a boy was walking on a trail” then I could give credit for the repetitive stem because it is different and I know that the student knows the word.
			* so if it seems like the student just chose a stem and threw in the target word, they do not get credit. If they chose a stem and each sentence is purposeful and correct, then a stem is fine.

S4 Practice Table (B)

These sentences are sample sentences selected from kindergarten children in Bilingual and Structured English Immersion classes. Normally, this table would reflect the responses given by just one child, but because this is for training purposes, it is important that the sentences selected reflect what the rater needs to know.

Word	Sentence	Score
School	school is for to do work and eat lunch	
Face	face has eyes	
Hop	hops can hop on the water	
Climb	climb	
Mittens	what	
Caterpillar	caterpillar can crawl and tickle our knees	
Born	a baby is born	
Feathers	I see a feather	
Woods	woods are from the	
Scarf	scarf is for when you cold and and uhm you are outside playing	
Munch	you munch the carrot	
Swooped	I like to eat soup	
Spring	a mi me gusta la primavera	
Crowd	all the people	
Squirm	squirm is when the squirm is wiggly	
Shelter	no response given	
Perch	perch is that you can look at it	
Trail	trail is a thing you can play	

Progression Chart	School	Face	Hop	Climb	Mittens	Caterpillar	Born	Feathers	Woods	Scarf	Munch	Swooped	Spring	Crowd	Squirm	Shelter	Perch	Trail
No Response *(if you checked one or more do not proceed and assign a 0 for that word)*																		
No Answer Given or Not in English																		
If any checked - Stop & assign (0) pts.																		
No Knowledge *(if you checked one or more do not proceed and assign a 1 for that word)*																		
Code-Switching																		
Incorrect Response																		
Repeated Target Word (and can't be rated)																		
Repetitive Stem Use																		
If any checked - Stop & assign (1) pt.																		
Some Knowledge *(if you checked two or more, please see if you can proceed to the next category)*																		
Shows Partial or Incomplete BUT Correct Knowledge																		
or Shows Complete & Correct Knowledge																		
Word Association (single or phrase)																		
Syntax errors BUT they do not hinder response																		
or No Syntax Errors																		
See if you can progress - if no, assign (2) pts.																		
K + Simple Sentence *(if you checked two or more, please see if you can proceed to the next category)*																		
Is there a subject & verb																		
Is there a subject & verb & object																		
Syntax errors BUT they do not hinder response																		
or No Syntax Errors																		
Complete Thought																		
Context is appropriate																		
See if you can progress - if no, assign (3) pts.																		

K+ Elaborate Sentences *(if you checked any here, then the score is a 4)*

may include prepositional phrases
or may include compound (subj., pred., or object)
or may include modifiers (adv & adj)
or may have many details
if any of these, assign (4) pts.

APPENDIX M

Project ELLA
STELLA
Story–retell Time for English Literacy and Language Acquisition

---

Little Rabbit's Journey

By: Beverly J.Irby/ Rafael Lara Alecio

Illustrated by Eva Vagretti Cockrille

Materials:

The Little Rabbit's Journey
Picture Word Cards

ESL Strategy: Interactive Read Aloud, Think Aloud

Vocabulary:

munch **boulder** **swooped**

Day 1

Introduce Vocabulary

(Point to the title.)

- Say Our story is called **The Little Rabbit's Journey.**
  (Point to the author's name.)
- Say The **authors** of the book are **Beverly J.Irby/ Rafael Lara Alecio**. Remember, the author writes the story.
- Say Would you like to be authors? What would you write about?
- Say Who can tell me what the **illustrator** does?
  (Point at the illustrator's name)
- Say The illustrator is an artist who makes pictures.
  (Point at title again.)
  Say Now, the title of the story is **The Little Rabbit's Journey.**.
- Say Do you know what a **journey** is**?**
- L1 Clarification: ¿Saben ustedes lo que es salir de viaje?
  (Wait for students to respond.)
- Talk about any personal journey you enjoyed and ask the students about their experiences during any particular journey.

- Say Today we are going to learn three new words. I want you to pay close attention because these are grown up words.
- (Show the picture card **munch**.)
- Say This is **munch**.
  (Read the sentence on the back of the card.)
- Say To **munch** is to chew food with a crunching sound.
- L1 Clarification: "To munch" quiere decir masticar con alegría.
  (Model answer using the stem, found on the back of the card. Students should answer in a complete sentence. If the student responds with a single word, make sure you model a complete sentence using the student's word and ask the child to repeat after you.
- (Model using the stem, Rabbits like to munch on…, found on the back of the card.
  Say Rabbits like to munch on… green plants.
  What do you think? Rabbits like to munch on…
  (Wait for students to respond. Students should answer in a complete sentence.)
  Say Let's pretend that we are munching.
  L1 Clarification: Vamos hacer como que estamos masticando con alegría.
  Say I like to munch carrots, I like to munch…

  (Show the picture card of **boulder**.)
- Say Who can tell me what this is?
  (Wait for students to respond)
- Say This is a picture of a **boulder.**
  L1 Clarification: Esta es la lámina de una roca.
  (Read the sentence on the back of the card.)
- Say **A boulder is a large stone in a stream.**
- LI Clarification Una roca es como una piedra grande en un riachuelo.
- Say Have you seen a **boulder before**? Where?
  (Wait for students to respond.)
  Model the answer using the stem, found on the back of the card. Students should answer in a complete sentence. If the student responds with a single word, make sure you model a complete sentence using the student's word and ask the child to repeat after you.
  (Model using the stem, A boulder looks like …, found on the back of the card.
  Say A boulder looks like…a giant. Your turn, A boulder looks like…

- (Show the picture card of **swooped**.)
- Say Who can tell me what this is?
  (Wait for students to respond)

- Say The students might respond with the name of the animal. When a bird like this one dives suddenly, the action is called swoop.
  This is a picture of a bird that **swooped.**
  (Read the sentence on the back of the card.)
- Say **Swoop is to dive or pounce suddenly like a hawk on its prey.**

  Model answer using the stem, found on the back of the card. Students should answer in a complete sentence. If the student responds with a single word, make sure you model a complete sentence using the student's word and ask the child to repeat after you.
  (Model using the stem, The eagle swooped …, found on the back of the card.
  Say The eagle swooped… like a hawk.
- Say Now it's your turn. …. The eagle swooped …
  (Wait for students to respond.)

- Say You all have done a wonderful job using the new words in complete sentences.

Activate and discuss background knowledge relating to the story

This is a book is about a rabbit who wanted to know what was on the other side of a mountain. It tells about the extra help needed to reach the other side and all the trouble he went through, just to find out that he didn't like what he saw on the other side of the mountain and decided to return to his place of origin.

Make connections to previous lessons or literature.

Introducing the Book

(Point to the book and say:)

- Say Looking at the cover of our book, who can tell me what the story might be about?
  (Wait for students to respond.)
  (Point to the rabbit on the cover.)

- Say Again, what is the name of this animal?
  (Wait for students to respond.)
  Make a topic web on the chalkboard or chart paper, write the word or draw a rabbit. What can you tell me about rabbits?
  Write down students' answers and review them when finished.

- (Show the cover of the book to the class.)
- Say Let's look at the cover of the book.
- Say As you can see, it is very colorful. Can you name some of the colors you see?
  (Wait for students to respond.)
- Say Does the cover give you a clue of what the story is going to be about?
  (Wait for students to respond.)
- Say Can you name a possible character of the story at this point?
  (Wait for students to respond.)
- Say What is this rabbit wearing?
  (Wait for students to respond.)
- Say What things would you put in a backpack?
- (A backpack.)

Day 2

Review vocabulary

- Say Remember we talked about three new words yesterday? Who can tell me what they were?
  (Wait for students to respond.)
- That's correct!
  (Show the picture card for **munch.)**
- Say The first word was **munch** and it means **to eat with happiness**.
-
  (Show the picture card for **boulder.)**
- Say The second word was **boulder** and it means**, a large rock in a stream.**
- (Show the picture card for **swooped.**
- Say The third word was **swooped, to dive or pounce suddenly like a hawk on its prey.**
- Let's repeat the words together. Ready?
- (Show the picture card for **munch**.) Say **Munch**.
  (Students should repeat with you.)
- (Show the picture card for **boulder.** Say **Boulder**.
  (Students should repeat with you.)
- (Show the picture card for **swooped**.)Say **Swooped**.
  (Students should repeat with you.)

Introduce the main characters

(Show the cover of the book.)

- (Point to the title.)
- Say Do you remember the title of the book?
  (Wait for students to respond.)
- Say Yes the title of the book is **The Little Rabbit's Journey.**
  (Wait for students to respond.)
- Say Looking at the cover of the book, can you tell me who one character in the story might be?
  (Wait for the students to respond.)
  Say How can you tell?
  (Wait for students to respond.)
- Who do you think might be some other characters in the book?
  (Wait for students to respond.)
- Let's read the story now and find out if there are other characters.

(READ the story with enthusiasm and expression. Stop where indicated and ask the following predictive and summative questions that will motivate students to recall story details. Wait for students to respond. Encourage the development of a dialogue stimulated by the questions.)

Begin reading story. Wait for students to respond after each question.

Page 2	• Where is the little brown rabbit? (Garden.) • What is he doing? (Munching lettuce.) • What kind of vegetable grows in this garden? (Point at the armadillo. Say:) Do you know the name of this animal? (Armadillo) • What is the armadillo doing?
Page 3	• What is the Little Rabbit looking at in the distance? (A mountain.) • How does the mountain look? (Very tall.) • What do you think is on the other side of the mountain? (Accepts students responses) • What is considered a "perfect place"?
Page 4	• What is the rabbit asking Mrs. Owl? • What is the rabbit wondering about? • What would you do if you wanted to find out what's on the other side of the mountain?
Page 5	• What was Mrs. Owl's advice to the rabbit? (Travel to the top of the mountain.)

	• What did Mr. Owl suggest to the little rabbit to take with him as he hops up the trail? • What decision does the little rabbit have to make?
Page 7	• Why did the little brown rabbit stop by the stream? (Drink water.) • What's the name of the animal in the stream? (Accept fish.) • How does the little brown rabbit feel? • Why is that so?
Page 8	• How is the salmon helping the little brown rabbit? • How do you think the little brown rabbit feels now?
Page 9	• Whom do you think will help the little rabbit and how?
Page 11	• Look at the little rabbit's face. Can you tell how he feels now? • Why do you think the little rabbit believes that the other side is the right place for him? • What is going to happen next? • Who is coming to help the little rabbit?
Page 13	• What is on the other side? (The city.) • Is the little brown rabbit happy now, why? • What is going to happen next? (Go back to the country.)
Page 14	• Name things the little rabbit saw in the city. • Do you think the little brown rabbit will stay? • What would you do if you were the little brown rabbit?
Page 15	• Name places the little rabbit went through on his way back home.
Page 16	• Which is the perfect place for the little rabbit? • Is the little rabbit happy now?
	• What did you like best about the story? • What surprised you the most about the story?

Tomorrow we will see how the story goes without interruptions.

Day 3

Story Review

Remember we talked about three new words before I read the story. Let's review them. They were:

- (Hold up the picture card for **munch**.)Say **Munch.**
- (Hold up the picture card for **boulder**.) Say **Boulder.**
- (Hold up the picture card for **swooped**.) Say **Swooped.**
- Say Now I am going to read the story again, and this time I want you to listen for the three words, munch, boulder, and swooped. When you hear me read the words **munch**, **boulder** and **swooped**, I want you to give a 'thumbs up' sign.
  (If needed, model 'thumbs up' for "Yes", until students' responses are firm.)
  (Begin reading story. Pause slightly after reading each of the three words to give students a chance to hear and put 'thumbs up'.)

Invite students to recall the title and author

(Point to the cover of the book.)

- Who remembers what the title of our book is?
  (Wait for students to respond. Prompt if necessary.)
- Yes! The title of our book is The Little Rabbit's Journey.
  (Wait for students to respond.)
- Who remembers what an author does?
  (Wait for students to respond. Prompt if necessary.)
- Yes! Authors write stories.
  (Point to author's name.)
- The name of the author is: (Point and read the name of the author and read the name.)
- Who remembers what the story was about?
  (Wait for students to respond. Prompt if necessary.)
- What would you have done if you were The Little Rabbit?
  (Wait for students to respond. Prompt if necessary.)

Encourage students to recall story characters.

- Now, who remembers what story characters are? (The people or animals in the story.)
- Who can name the character from this story?
  (Wait for students to name the character.)

- Did you like the character? Do you remember what the character did in the story?
  (Wait for students to respond.)
  (Accept reasonable responses.)

- Can you recall times when you needed help?
  (Wait for students to respond.)

- How did you feel when you needed help?
  (Wait for students to respond.)

- Did you ask someone to help you?
  (Wait for students to respond.)

- What did you learn from the story?
  (Wait for students to respond.)

- What happened to the rabbit at the end?
  (Wait for students to respond.)

Story Critique

Invite children to share their literary opinions in a risk-free setting.

- Now we are going to talk about what we liked about this story and what we didn't like about this story. We are going to be story critics.
- This is how we are going to do it. I want you to put your thumbs up (put thumbs up) to tell me, "Yes, I liked that", or put your thumbs down (put thumbs down) to tell me, "No, I didn't like that."
  (Model strategy until students' response is firm.)
- O.K. ready? Do you like the title? Put your thumbs up for "Yes" and thumbs down for "No". (Participate with students.) Good!
  (Ask a student to count the number of thumbs up.)
  (Continue this process with the rest of the questions.)
- Do you like the pictures in the story?
- Do you like the characters in the story?
- Do you like the illustrations?

On your story chart, place or draw a peanut under each section that receives the most votes. As children develop more fluency, their verbal participation will increase.

Day 4

Story Review

Review the story vocabulary as it relates to the story.

- Does anyone remember our three vocabulary words?
  (Wait for students to respond)
- Yes, they were **munch, boulder and swooped.**
- (Display the three picture cards.)
- I want you to point to the picture that matches the word I say and repeat the word.
- (Children should point to the picture of munch and say the word **munch**. Correct or redefine if needed**.)**
- Do you remember how the author used the word **munch**?
- (Wait for students to respond)
- That's right!
- (Repeat with the words **boulder** and **swooped**.)

Invite students to recall the title, author, and characters of the story.

- (Point to title of the story.)
- What is the title of our story?
- (Wait for students to respond.)
- Yes, the title of our story is **The Little Rabbit's Journey**
- And the author is (Point to the author's name and read.)
- Who are the characters in the story?
- (Wait for students to name the only character, the caterpillar.)

Interactive Group Retelling

Reread the story again.

1. Have a picture of a rabbit or stuffed rabbit.
2. Have children sit in a circle.
3. Begin a round-robin story with children in which each storyteller will make up a story of what would happen to The Little Rabbit if he came to Houston.
4. Give each child a maximum of two minutes for his or her section of the story.

Day 5

Reread the story.

Make a Trail Mix Treat for the students. Tell the students they are going to prepare a treat for The Little Rabbit's Journey. Separate the zip lock bags and give one to each student.

A. Mix in a Bowl:
1. Raisins
2. Peanuts
3. Sunflower seeds
4. M & M's

Give the directions in steps.

Or

Make a carrot salad for the class. Rabbits eat carrots and we do too.

B. Make a Carrot Salad.
1. Carrots – grind them
2. Raisins
3. Sunflower Seeds
4. Crushed Pineapple or tidbits
5. Mayonnaise to taste

Stir in a bowl and serve.

- Show students the book, The Little Rabbit's Journey.
- (Ask :) Do you remember what the rabbit ate on the garden at the beginning of the story?

Ask the students to draw the picture of the fruits as you guide them day by day.

A Learning center activity:
Have boxes to construct a city, students can pretend The Little Rabbit arrived at the city and discovered…

VITA

Name:	Miranda Fernande Walichowski
Address:	107 J Harrington Tower, TAMU MS 4225, College Station, TX 77843-4225
Email Address:	m-walichowski@tamu.edu
Education:	B.S., Maritime Business Administration, Texas A&M University at Galveston, 1997
	M.Ed., Hispanic Bilingual Education, Texas A&M University, 2002
	Ph.D., Educational Psychology, Texas A&M University, 2009