

**FACTORS AFFECTING NORMING: A DEVELOPMENTAL STUDY OF ORAL
LANGUAGE MEASURES IN SPANISH-SPEAKING
ENGLISH LANGUAGE LEARNERS**

A Dissertation
Submitted
to the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

By
Patricia A. Swasey Washington
January, 2010

©

Copyright

2010

by

Patricia A. Swasey Washington

ABSTRACT

Title: FACTORS AFFECTING NORMING: A DEVELOPMENTAL STUDY OF
ORAL LANGUAGE MEASURES IN SPANISH-SPEAKING ENGLISH LANGUAGE
LEARNERS

Candidate's Name: Patricia A. Swasey Washington

Degree: Doctor of Philosophy

Temple University, 2010

Doctoral Advisory Committee Chair: Aquiles Iglesias, Ph.D.

The Latino population of which English Language Learners (ELLs) is a subset, has demonstrated substantial growth in recent years (U.S Census Bureau, 2008), highlighting the need for normative information regarding their language skills. However, requisite to obtaining normative information is determining appropriate norming methods. The principal purpose of the present study was to ascertain appropriate norming procedures for the language variables: Mean Length of Utterance (MLU), Number of Different Words (NDW) and Words Per Minute (WPM) in English and Spanish narratives of Spanish-speaking ELLs. The issues were 1) whether age or grade norms should be used as an index of language development, 2) whether cross-sectional or longitudinal data should be utilized, and 3) whether the inclusion or exclusion of children with missing data or grade repeats affects the language measures. It was hypothesized that due to the syntactic and lexical differences that are present across languages, there would be a different developmental schedule of development for the language variables in the English and Spanish of ELLs. Participants were typically developing kindergarten to

second grade Spanish speaking ELLs enrolled in transitional bilingual programs. A total of 605 children comprised the cross-sectional dataset and a total of 679 children were included in the longitudinal dataset. From these initial datasets, additional datasets were created to provide separate age and grade groups (for the cross-sectional and longitudinal datasets) as well as three different longitudinal datasets. Narratives in English and Spanish were elicited from each child using a story retell procedure. Analyses were carried out using Multivariate Analysis of Variance (MANOVA), Univariate Analysis of Variance (ANOVA) and Repeated Measures Analysis of Variance procedures. Results of both cross-sectional and longitudinal analyses indicated that age and grade are comparable indices of time for studying MLU, NDW, and WPM. Results also indicated that longitudinal data is superior to cross-sectional data for examining the language variables and that including or excluding subjects with missing data or grade repeaters does not significantly affect MLU, NDW, and WPM scores. Additionally, results confirm the findings from the research literature that MLU, NDW, and WPM are valid variables for studying narrative development.

ACKNOWLEDGMENTS

First, I thank God for His Son, Jesus Christ, and for all that He has done and continues to do in my life.

I extend sincere gratitude to Dr. Aquiles Iglesias for taking me under his tutelage and guiding me in the path of scholarship. I appreciate his patience, kindness, encouragement, as well as his commitment to rigor and precision in research. He has been an exceptional mentor. I have always had great admiration for his expertise in the area of bilingual language development. I thank him also for providing the funding that made the goal of completing my doctoral studies achievable.

I am very grateful to Dr. Brian Goldstein for his encouragement to pursue my doctorate, a plan that I had for a long time, but whose implementation had become elusive. It was that initial research venture with him, I as a speech-language pathologist in the field and he as a professor at Temple University, that the desire to pursue my doctorate was rekindled. Finally, I saw it as an immediate, and not a too lofty or distant goal. I thank him also for his guidance and superior teaching during my years at Temple.

Thanks to Dr. Nadine Martin for serving on my committee and for her encouragement.

Thanks to Dr. Leah Fabiano-Smith for graciously accepting to be my external reader, and for her encouragement.

Thanks to Raúl Rojas for his encouragement, insightful comments, and assistance with questions related to the BLLP database.

Thanks to David Ford and the staff at the Social Sciences Data Library for providing me with statistical support.

Thanks to Dr. Aneta Pavlenko for her encouragement and her intellectually motivating classes in second language acquisition.

Thanks to all my family and friends. Thanks to Samuel (my husband) for agreeing with my decision to pursue the degree. Special thanks to Ephraim, Jared, and Manasseh (my sons) for their love, understanding and encouragement. Thanks to Guillemette and Robert Serlin for their encouragement. I especially thank Bishop Omega Shelton for his inspiring sermons that have helped to nurture my faith and spirit.

For Ephraim, Jared, and Manasseh, for their love, patience and understanding throughout my program of study.

For my mother and father (Monica and Patrick Swasey) for their love, prayers, support and encouragement.

TABLE OF CONTENTS

| | Page |
|---|------|
| ABSTRACT | iii |
| ACKNOWLEDGMENTS | v |
| DEDICATION | vi |
| LIST OF TABLES | x |
| CHAPTER | |
| 1. LITERATURE REVIEW | 1 |
| 1.1 Norming Of Standardized Tests And Language Sample Norms..... | 3 |
| 1.1.1 Norming Of English Standardized Tests | 5 |
| 1.1.2 Norming Of Spanish Standardized Tests..... | 7 |
| 1.1.3 Language Sample Norms..... | 11 |
| 1.2 Language Measures Based On Narratives..... | 12 |
| 1.2.1 Mean Length Of Utterance | 21 |
| 1.2.2 Number Of Different Words..... | 22 |
| 1.2.3 Words Per Minute..... | 24 |
| 1.3 Research Questions..... | 25 |
| 1.3.1 Norming By Age Or Grade?..... | 25 |
| 1.3.2 Cross-Sectional Or Longitudinal Data? | 30 |
| 1.3.3 Does Inclusion Or Exclusion Of Children With Missing Data Or Grade Repeats Affect Language Norms?..... | 33 |
| 1.4 Summary | 33 |

| | |
|---|-----|
| 2. METHODOLOGY | 37 |
| 2.1 Participants..... | 37 |
| 2.1.1 Cross-Sectional Sample | 38 |
| 2.1.2 Longitudinal Datasets | 40 |
| 2.2 Procedure | 44 |
| 2.2.1 Transcription And Coding | 45 |
| 2.2.2 Reliability..... | 47 |
| 2.2.3 Analysis Procedure | 48 |
| 3. RESULTS | 50 |
| 3.1 Cross-sectional Dataset..... | 50 |
| 3.2 Longitudinal Dataset..... | 66 |
| 3.3 Comparison Across Longitudinal Datasets..... | 84 |
| 4. DISCUSSION | 93 |
| 4.1 Age Or Grade As An Index Of Time?..... | 96 |
| 4.2 Cross-Sectional Or Longitudinal?..... | 101 |
| 4.3 Inclusion Or Exclusion Of Participants With Missing Data Or Grade Repeaters? | 104 |
| 4.4 The Language Measures, MLU, NDW, And WPM | 107 |
| 4.5 Study Limitations..... | 111 |
| 4.6 Future Research | 111 |
| REFERENCES CITED | 114 |
| APPENDICES | |
| A. MISSING DATA PATTERNS | 127 |
| B. BILINGUAL SPANISH/ENGLISH STORY RETELL STORY SCRIPTS | 129 |
| C. CALCULATION OF EFFECT SIZE | 132 |

LIST OF TABLES

| Table | | Page |
|--|--|------|
| 1. English Language Tests and Normative Sample Description | | 6 |
| 2. Spanish Language Tests and Normative Sample Description | | 10 |
| 3. Number and percentage of children who produced narratives in both English and Spanish, English only, and Spanish only at each data collection point (Longitudinal Dataset I) | | 42 |
| 4. Number and percentage of children who produced narratives in both English and Spanish, English only, and Spanish only at each data collection point (Longitudinal Dataset II) | | 43 |
| 5. Schedule of books for data collection..... | | 45 |
| 6. Mean Scores of EMLU, ENDW, and EWPM as a Function of Age (Cross- sectional Dataset) | | 53 |
| 7. Mean Scores of EMLU, ENDW, and EWPM as a Function of Grade (Cross-sectional Dataset) | | 54 |
| 8. Univariate ANOVA Results for EMLU, ENDW, and EWPM as a Function of Age (Cross- Sectional Dataset) | | 55 |
| 9. Univariate ANOVA Results for EMLU, ENDW, and EWPM as a Function of Grade (Cross- Sectional Dataset) | | 56 |
| 10. Post Hoc Tests for EMLU, ENDW and EWPM as a Function of Age (Cross-sectional Dataset) | | 57 |
| 11. Post Hoc Tests for EMLU, ENDW and EWPM as a Function of Grade (Cross-sectional Dataset) | | 58 |
| 12. Mean Scores of SMLU, SNDW, and SWPM as a Function of Age (Cross- sectional Dataset) | | 61 |
| 13. Mean Scores of SMLU, SNDW, and SWPM as a Function of Grade (Cross- sectional Dataset) | | 61 |
| 14. Univariate ANOVA Results for SMLU, SNDW, and SWPM as a Function of Age (Cross-sectional Dataset) | | 62 |
| 15. Univariate ANOVA Results for SMLU, SNDW, and SWPM as a Function of Grade (Cross- sectional Dataset) | | 63 |

| | | |
|-----|--|----|
| 16. | Post Hoc Tests for SMLU, SNDW and SWPM as a Function of Age (Cross- sectional Dataset) | 64 |
| 17. | Post Hoc Tests for SMLU, SNDW and SWPM as a Function of Grade (Cross- sectional Dataset) | 65 |
| 18. | Longitudinal Dataset III: English Language Variables by Age (N=72)..... | 68 |
| 19. | Longitudinal Dataset III: English Language Variables by Grade (N=74)..... | 69 |
| 20. | Repeated Measures ANOVA Results for EMLU as a Function of Age (Longitudinal Dataset III) | 70 |
| 21. | Repeated Measures ANOVA Results for ENDW as a Function of Age (Longitudinal Dataset III) | 70 |
| 22. | Repeated Measures ANOVA Results for EWPM as a Function of Age (Longitudinal Dataset III) | 70 |
| 23. | Least Significant Difference Comparisons for EMLU as a Function of Age (Longitudinal Dataset III) | 71 |
| 24. | Least Significant Difference Comparisons for ENDW as a Function of Age (Longitudinal Dataset III) | 71 |
| 25. | Least Significant Difference Comparisons for EWPM as a Function of Age (Longitudinal Dataset III) | 72 |
| 26. | Repeated Measures ANOVA Results for EMLU as a Function of Grade (Longitudinal Dataset III) | 73 |
| 27. | Repeated Measures ANOVA Results for ENDW as a Function of Grade (Longitudinal Dataset III) | 73 |
| 28. | Repeated Measures ANOVA Results for EWPM as a Function of Grade (Longitudinal Dataset III) | 73 |
| 29. | Least Significant Difference Comparisons for EMLU as a Function of Grade (Longitudinal Dataset III) | 74 |
| 30. | Least Significant Difference Comparisons for ENDW as a Function of Grade (Longitudinal Dataset III) | 74 |
| 31. | Least Significant Difference Comparisons for EWPM as a Function Of Grade (Longitudinal Dataset I) | 75 |
| 32. | Longitudinal Dataset III: Spanish Language Variables by Age (N=157) | 76 |
| 33. | Longitudinal Dataset III: Spanish Language Variables by Grade (N=163) | 77 |

| | | |
|-----|--|----|
| 34. | Repeated Measures ANOVA Results for Spanish MLU as a Function of Age (Longitudinal Dataset III) | 78 |
| 35. | Repeated Measures ANOVA Results for Spanish NDW as a Function of Age (Longitudinal Dataset III) | 78 |
| 36. | Repeated Measures ANOVA Results for Spanish WPM as a Function of Age (Longitudinal Dataset III) | 78 |
| 37. | Least Significant Difference Comparisons for Spanish MLU as a Function of Age (Longitudinal Dataset III) | 79 |
| 38. | Least Significant Difference Comparisons for Spanish NDW as a Function of Age (Longitudinal Dataset III) | 79 |
| 39. | Least Significant Difference Comparisons for Spanish WPM as a Function of Age (Longitudinal Dataset III) | 80 |
| 40. | Repeated Measures ANOVA Results for SMLU as a Function of Grade (Longitudinal Dataset I)..... | 81 |
| 41. | Repeated Measures ANOVA Results for SNDW as a Function of Grade (Longitudinal Dataset III) | 81 |
| 42. | Repeated Measures ANOVA Results for SWPM as a Function of Grade (Longitudinal Dataset III) | 81 |
| 43. | Least Significant Difference Comparisons for SMLU as a Function of Grade (Longitudinal Dataset III) | 82 |
| 44. | Least Significant Difference Comparisons for SNDW as a Function of Grade (Longitudinal Dataset III) | 82 |
| 45. | Least Significant Difference Comparisons for SWPM as a Function of Grade (Longitudinal Dataset III) | 83 |
| 46. | Dataset Comparisons for Kindergarten, First Grade, and Second Grade for EMLU, ENDW, EWPM and SMLU, SNDW, and SWPM..... | 87 |
| 47. | Longitudinal Dataset Comparisons for EMLU, ENDW, and EWPM: First Grade..... | 88 |
| 48. | Longitudinal Dataset Comparisons for EMLU, ENDW, and EWPM: Second Grade..... | 89 |
| 49. | Post Hoc Tests for First grade Longitudinal Datasets: Comparisons by EMLU, ENDW and EWPM | 90 |

| | |
|--|----|
| 50. Post Hoc Tests for Second Grade Longitudinal Datasets: Comparisons by EMLU, ENDW and EWPM | 91 |
|--|----|

CHAPTER 1

LITERATURE REVIEW

The Hispanic population, individuals who originated from Mexico, Puerto Rico, Cuba, Spanish-speaking Central and South American countries, and other Spanish cultures, represents more than half of the overall population growth in the United States since 2000 (U.S. Census Bureau, 2008). By 2007 this group comprised 15.1 percent of the U.S. population and is currently the largest ethnic minority group in the country (Pew Hispanic Center, 2008). Moreover, due to rapid growth, the Latino population is projected to become 30 percent of the U.S. population by 2050 (U.S. Census Bureau, 2008). A subset of the Hispanic population is children whose primary language is one other than English. These children, referred to as English Language Learners (ELLs), include over five million kindergarten to 12th grade students, representing 10.5 percent of the U.S. student population in 2004-2005 (NCELA, 2006). Based on data reported by states for the 2001-2002 school year, about 79 percent of all ELLs were speakers of Spanish (NCELA, 2008). Despite the growing presence, as well as projected increase of Spanish-Speaking ELLs (Pew Hispanic Center, 2008; U.S. Census Bureau, 2008), there is insufficient normative information regarding their language skills. The limited amount of available data is primarily based on either small scale or cross-sectional studies (such as normative data used to create traditional standardized tests). Obtaining more comprehensive normative data is important in order to determine the accurate developmental sequence of language skills and could eventually lead to the development of more valid evaluation instruments and sounder intervention goals. For ELLs, it is

imperative that norms are available in both of their languages and based on a sample representative of the population.

Due to concerns about test item bias of many of the commercially available tests, alternative assessment approaches have been recommended (Dunn, Flax, Sliwinski, & Aram, 1996; Gavin, Klee & Membrino, 1993; Scott & Windsor, 2000). An alternative or supplemental assessment approach that has been proposed is the use of narrative language samples (Hughes, McGillivray, & Schmidek, 1997). Language sampling has been demonstrated to be more relevant across cultures and is capable of eliciting a large range of language skills (e.g., Leadholm & Miller, 1992; MacLachlan & Chapman, 1988). Unfortunately, normative data on language acquisition based on language samples have been lacking (Muñoz, Gillam, Peña, & Gulley-Faehnle, 2003).

Essential to the creation of adequate normative data on language acquisition are several issues that have received little or no attention in the language developmental literature. The issues that require further study include whether age or grade is the best index of time in studies examining language development, whether cross-sectional or longitudinal data should be used in the construction of language acquisition norms, and whether the inclusion or exclusion of children with missing data or grade repeaters affects the language acquisition norms.

In order to obtain appropriate language acquisition norms for ELLs, it is critical to examine whether there are differences in the norms depending on how they are collected and constructed. This study will investigate different methods of collecting and constructing language norms for the ELL population. It is expected

that results from this study will assist in the development of future norms on language development of ELLs. The existing literature on present practices of norming standardized tests and measures obtained from language samples will be discussed first. This will be followed by a review of existing studies on language sampling, language outcome variables, the use of grade and age as time indices of development, and finally, differences between norms obtained with cross-sectional and longitudinal data.

Norming of Standardized Tests and Language Sample Norms

A major issue with standardized tests is the lack of adequate norms (Erickson & Iglesias, 1986; Langdon, 1992). This has been a problem for English and Spanish tests used with Spanish-speaking ELLs. Some tests have no available norms, while others have small norming samples (Laing & Kamhi, 2003). Additionally, some tests do not take into consideration the heterogeneity of the ELL population.

The specific purpose of a standardized test will determine the type of normative sample that is required (Bedore & Peña, 2008; Merrell & Plante, 1997; Plante & Vance, 1994, 1995). When selecting a norming sample for the purpose of differentiating between language delay and typical development, it is necessary to choose individuals that are representative of the typically-developing population (Bedore & Peña, 2008; Peña, Spaulding & Plante, 2006). As noted by Peña et al. (2006), tests that include typically developing and language disordered children under-identify cases of language impairment because they create an overlap between normal and impaired samples. Thus, norms should be developed based on typically

developing children. In addition, it is important that the norming sample have representation from individuals with the linguistic and cultural backgrounds of the individuals with whom the test will be used (Bedore & Peña, 2008; Saenz & Huer, 2003).

When norming tests for ELL children it is important to take into consideration the similarities and differences that exist across the two languages. These variations are due to structural and cultural differences across the languages and the interaction between the two languages. For example, similarities have been reported in the number of words at certain points in development (Holowka, Brosseau-Lapr e, & Pettito, 2002; Pearson, Fern andez & Oller, 1993, 1995). However, the lexicons of ELLs often contain identical and unique vocabulary items across the two languages (e.g., Pearson & Fern andez, 1994; Pe a, Bedore, & Zlatic-Giunta, 2002). In their study, Pe a et al. (2002) reported that while bilingual Spanish-English children named a comparable number of words in a “category generation task,” they provided distinct items in each language about 70 percent of the time. The items produced for each language was highly associated with particular activities linked to certain language contexts. In addition to demonstrating differing skill levels for both of their languages, ELLs can possibly exhibit lower skill levels in certain aspects of their first language (Anderson, 1999) as the second language develops. These findings support the notion that measures selected for norming should be language neutral and, perhaps, different norms should be created for each of the languages.

Norming of English Standardized Tests

Standardized English assessment tools that have been traditionally used with ELLs have tended not to include culturally and linguistically diverse children (Laing & Kamhi, 2003) in their normative groups. For those tests that included these children in the standardization sample, however, there is concern that often the samples are not adequately representative of the populations being tested (Laing & Kamhi, 2003). Recent tests, for example, the Preschool Language Scale-4 (PLS-4 English; Zimmerman, Steiner & Pond, 2002), and the Clinical Evaluation of Language Fundamentals-4 (English edition) (CELF-4 English; Semel, Wiig, & Secord, 2003) have included English-speaking Hispanic children based on the representation of Hispanics in the national census. Table 1 presents commonly used language tests and information on the standardization sample.

As can be seen in this table, recent standardized test norms have included Hispanics, as represented in the US Census. Hispanics comprised 15.1 percent of the United States population for 2007 (U S Census Bureau, 2008). However, the percentage provided by the US Census might be at variance with the reports of Hispanics in the educational system. In the educational system, Hispanic children ages 5-17 years old make up about 20% of the Hispanic population (US Census Bureau, 2007), and test developers tend to use the overall population statistics of 15.1 which would result in undersampling of the Hispanic children for test standardization. Further, it is unclear how many of these Hispanic children are also Spanish speakers and the extent to which they are proficient English speakers. Census data suggest that

13 percent of the US population five to 17 years of age in 2000 were speakers of Spanish (US Census, 2000) and that 38% percent of these children did not speak English “very well.” It is impossible to ascertain how many of these children who do not speak “English well” are part of any norming sample.

Table 1. English Language Tests and Normative Sample Description

| Tests | Sample Size & Age Range | Description Of Normative Sample |
|-----------------------|------------------------------------|---|
| CELF-4 English (2003) | 2,650 5- to 21-year-olds | English-speaking Based on 2000 U.S Census, including students in special education. |
| EOWPVT-English (2000) | 2,327 2;0 to 18;11 | Primarily English-speaking. Representative of U.S. Census |
| PLS-4 English (2002) | 1,564 2 days to 6;11 | English-speaking Based on 2000 U.S Census for children birth-6 years, including those with disabilities |
| ROWPVT-English (2000) | 2,327 2;0 to 18;11 | Primarily English-speaking. Representative of U.S. Census |

It is important to note, however, that including culturally and linguistically diverse populations might just merely alter the mean of the population and not improve on the validity of the instrument. Often culturally and linguistically diverse children, who comprise a small proportion of the standardization population for language measures, will perform lower than the average of the population (Heaton & Marcotte, 2000). As a result, it is impossible to ascertain their true linguistic performance because the indices for comparison are not ideal. Although it is good practice to include a representative sample of culturally and linguistically diverse individuals in norming of standardized tests, the best solution to this dilemma is to create assessment tools for specific populations of culturally and linguistically diverse individuals (Liang & Kamhi, 2003).

The degree of English proficiency of Hispanics participating in the norming sample has an effect on the resulting language norms. Therefore, if individuals included in the norming sample for a test have low average English proficiency, the comparison scores will be low. Conversely, if individuals in the norming sample have high English language proficiency, test scores used for comparison purposes will be high. Due to this concern, it is necessary that test instruments adequately determine and describe the language skills of populations included in their norming samples. The criteria for determining language proficiency for inclusion of individuals in the standardization population can be specified by researchers of the tests. For example, requirements of some English tests that include Spanish-speaking individuals in the standardization population, is that English must be their primary language (e.g., CELF-4 English; Semel et al., 2003). However, their actual English language proficiency might not be ascertained due to non-objective means of acquiring this information (e.g., questionnaires to parents, examinees, or to school personnel).

Norming of Spanish Standardized Tests

A variety of Spanish tests have been normed for use with Spanish-speaking ELLs. Although some of these instruments have extensive normative information based on representative samples of Spanish-speaking children in the United States, others fall short of this requirement (Langdon, 1992). For example, some tests include only norms for monolingual Spanish speakers while others do not take into account ELLs who might be in different stages of development of Spanish (Langdon, 1992).

The language skills of ELLs are frequently different from those of monolingual children of either of their target languages (e.g., Goldstein, 2004). Therefore, using monolingual norms with Spanish-speaking ELLs might invalidate test results. For example, the Spanish version of the Peabody Picture Vocabulary Test (Test de Vocabulario en Imágenes Peabody (TVIP; Dunn, Lugo, Padilla & Dunn, 1986)) was normed on monolingual Spanish-speaking children in Mexico and Puerto Rico, making the overall validity of this measure questionable when applied to non-monolingual Spanish-speaking populations. Results of pilot studies conducted by Dunn et al. (1986) showed that bilingual children scored at approximately one standard deviation below the mean compared to monolingual Spanish-speaking children from Mexico and Puerto Rico.

Recent assessment instruments have been designed to be more sensitive to Spanish-speaking ELLs by involving sizeable normative samples of the bilingual Spanish-English population in the United States. Norming samples for these tests are representative of the Hispanic population in that they include ELLs based on U.S. Census data for Hispanic households organized according to variables such as parental education and geographical region. Such instruments include the PLS-4 Spanish (Zimmerman, Steiner & Pond, 2002), the CELF-3 Spanish (Semel, Wiig, & Secord, 1997), the Clinical Evaluation of Language Fundamentals -4 Spanish (CELF-4 Spanish; Wiig, Secord, & Semel, 2005), the EWOPVT-SBE (Brownell, 2001), the ROWPVT-SBE (Brownell, 2001) and the Test of Early Language Development-Third Edition: Spanish (TELD-3: Spanish; Ramos & Ramos, 2007). Table 2 presents information on frequently used Spanish language tests and a description of their

normative sample. The use of U.S Census data might also be inappropriate for selecting the norming sample for Spanish tests since the children in the various Hispanic subgroups are not equally represented. For example, 1993 data indicate that Mexicans constitute 64% of the total Hispanic population. In comparison, Cubans represent 4.7 of the population. However, the Cuban population had the highest median age (43.6 years), while Mexicans had the lowest median age (24.6 years). Therefore, if these statistics are used to determine the norming population for tests used with Hispanics, there will be an oversampling of Cubans because this group has a relatively smaller percent of young people.

As with tests in English, in the process of obtaining a sample for norming Spanish tests, the Spanish language proficiency of the participants is often not adequately determined. Language requirements for inclusion in the standardization sample are varied, making the populations used for standardization across tests not comparable. For example, for the Preschool Language Scale-4 Spanish (PLS-4 Spanish; Zimmerman et al., 2002), individuals are required to speak Spanish fluently, for the Clinical Evaluation of Language Fundamentals- Third Edition Spanish (CELF-3 Spanish; Wiig et al., 2005), participants must understand and speak Spanish fluently, and for the Expressive One Word Picture Vocabulary Test- Spanish Bilingual Edition (EOWPVT-SBLE; Brownell, 2001) and the Receptive One Word Picture Vocabulary Test – Spanish Bilingual Edition (ROWPVT-SBLE; Brownell, 2001) they only need to speak “at least some Spanish” (Brownell, 2001, p. 61). For some tests, language proficiency was determined by test examiners based on their judgment of the individual’s language (EOWPVT-SBLE and ROWPVT-SBLE-

Brownell, 2001; PLS-4 Spanish –Zimmerman et al., 2002), on responses to questionnaires given to the examinees (EOWPVT-SBLE and ROWPVT-SBLE- Brownell, 2001; CELF-3 Spanish- Wiig et al., 2005), or on report of parents or school districts (CELF-3 Spanish; Wiig et al., 2005). Subject selection based on Census data also exacerbates the problem of obtaining a representative normative sample because Census data do not identify the particular numbers of children with specific language proficiency in Spanish that could be later used for deriving normative samples for tests. A more complete and accurate method of ascertaining the language skills of ELLs is undoubtedly the narrative. It gives a varied sample of language performance and captures the heterogeneous nature of ELL language use.

Table 2. Spanish Language Tests and Normative Sample Description

| Tests | Sample Size & Age Range | Description Of Normative Sample |
|-------------------------|--------------------------------------|---|
| CELF-4 Spanish (2006) | More than 1,100 5 to 21-year-olds | Normed on representative sample of U.S. Spanish speakers, including students receiving school services (e.g., English as a second language, speech-language services) |
| EOWPVT-Bilingual (2001) | 1,050 4;4 to 12;11 | Bilingual (Spanish-English) individuals Representative of U.S. Hispanic population. |
| PLS-4 Spanish (2002) | 1,188 2 days to 6;11 | Representative sampling of Spanish-speaking children based on 2000 U.S. census information for Hispanics. Includes children with language delays. |
| ROWPVT-Bilingual (2001) | More than 1,000 4;0 to 12;11 | Bilingual (Spanish-English) individuals Representative of U.S. Hispanic population. |
| TELD-3 Spanish (2007) | 1,441 2;0 to 7;11. | Children living in Chile, Costa Rica, Mexico, Spain, and the United States. |

Language Sample Norms

While narratives have been attested to be a suitable alternative or complement to standardized tests, the norms currently available for them are lacking. Normative information based on narrative language sample analysis of Spanish-speaking ELLs usually involve cross-sectional data (e.g., Miller, Heilmann, Nockerts, Iglesias, Fabiano, & Francis, 2006; Muñoz et al., 2003), small sample sizes, and evaluation primarily of only one language (e.g., Muñoz et al., 2003). The only available comprehensive language sample norms for ELLs in both of their languages are those provided by the Systematic Analysis of Language Transcripts (SALT) database (Miller & Iglesias, 2007). The “Bilingual Spanish/English Story Retell reference databases” currently include narratives of more than 2,000 kindergarten to third grade (ages 5;0 to 9;9) native Spanish-speaking ELLs from educational programs in urban and border areas of Texas and urban areas of California (SALT, 2009). A variety of standard measures are available, including Mean length of utterance (MLU), Number of different words (NDW), and Words per minute (WPM). Following the collection of language samples, transcriptions are facilitated through SALT’s editor. Subsequently, various standard reports can be generated and a child’s performance can be compared to the performance of age- or grade- matched peers in the database. Obviously, the size of the comparison group to which the child’s performance is compared would vary depending on the number of variables specified for the comparison group. Using a first grader, age 7;9 as an example, the comparison group consists of 848 children when an age match of plus or minus six months is used. In contrast, the comparison group is 298 when a grade match is selected. If both age and

grade are specified, then the comparison group is further reduced to 97 children. Although these norms are useful, they are based on cross-sectional data. Cross-sectional data provide a mere snapshot of performance; they do not have the ability to detail development over time. Standardized tests and language sample norms used with Spanish-speaking ELLs have been based on cross-sectional data.

The research literature suggests that longitudinal data be used to examine language development (Genesee & Nicoladis, 2007; Szaflarski, Schmithorst, Altaye, Byars, Ret, Plante, & Holland, 2006). The need and initiatives for collecting longitudinal data are supported by a host of government grants (e.g., U.S. Department of Education, 2009). Moreover, a variety of statistical methods are available for evaluating longitudinal data such as latent growth curve modeling (e.g., Duncan, Duncan & Strycker, 2006) and hierarchical linear modeling (Raudenbush & Bryk, 2002). There are some disadvantages of using longitudinal data, with the biggest disadvantage being the time it takes to collect the data. Prior to moving away from using cross-sectional data for norming language skills, it is important to examine the extent to which normative data obtained using cross-sectional data would differ from norms based on longitudinal data.

Language Measures Based on Narratives

There is consensus in the field of Communication Sciences and Disorders that appropriate evaluation of children who are culturally and linguistically diverse should include more than standardized assessments (e.g., Brice, 2002; Goldstein, 2004; Laing & Kamhi, 2003; Langdon, 1992; Langdon, 2008; Washington, 1996). One

alternative approach often recommended is the use of language sampling, specifically narratives. The use of measures derived from oral language samples, combined with parent reporting are the most accurate way for determining which Spanish-speaking ELLs present with a language delay (Gutierrez-Clellen, Restrepo, Bedore, Peña, & Anderson, 2000). This approach to assessment has also been successful in the evaluation of English-speaking children (e.g., Dunn, Flax, Sliwinski, & Aram, 1996; Gavin, Klee, & Membrino, 1993; Scott & Windsor, 2000). Language sampling (Leadholm & Miller, 1992; Miller et al., 2006) has been accepted as a valid method for assessing children's language proficiency (Aram, Morris, & Hall, 1993; Hewitt, Hammer, Yont, & Tomblin, 2005) and should be included in the language evaluation process of all culturally and linguistically diverse children (Battle, 2002; Brice & Montgomery, 1996; Cheng, 1991; Kayser & Restrepo, 1995; Mattes & Omark, 1991).

Narrative language sampling, particularly retelling a narrative after a model (retell condition), is ideal for assessing the linguistic skills of Spanish-speaking ELLs because it allows for the sampling of a large range of linguistic skills. Importantly, the analyses of specific measures derived from narratives are language neutral and permit cross language comparisons.

The narrative is regarded as an effective means of assessing expressive language skills of children from all cultures since all cultures expose their children to oral storytelling (Hughes et al., 1997). Moreover, narrative skill is a positive predictor of linguistic and academic skills of English-speaking mainstream (Bishop & Edmundson, 1987) and culturally and linguistically diverse children (e.g., Fazio, Naremore, & Connell, 1986). In a longitudinal study by Bishop and Edmundson

(1987), 87 children in England with language impairment were evaluated at the ages of four, four and a half, and five and a half using several language measures. Resolution of the language impairment (good or poor outcomes) was able to be predicted with 90% accuracy based on measures taken at four years of age. The best predictor was story retelling with pictorial support. Similar results were found in other studies. For example, Fazio et al. (1986) conducted a three-year longitudinal study with 34 kindergarten through second grade children “from poverty,” in order to differentiate between children with specific language impairment and those at the low normal range. They found that the best predictor at kindergarten of academic level for 15 of the children who received remedial assistance was story retelling.

Similar to English-speaking children, oral narratives from ELL children can be used to assess language development. Heilmann, Miller, Iglesias, Fabiano-Smith, Nockerts, and Andriacchi (2008) found high levels of transcription accuracy for English and Spanish across transcribers using 40 narratives of ELLs. Test-retest reliability using 241 transcripts of ELLs (115 in English and 126 in Spanish) showed high correlations between the two time periods for four narrative measures (MLU, NDW, Number of total words, and WPM). These findings suggest that narrative information can be accurately transcribed and that the narrative measures are reliable across time. In another study, from their examination of narratives of more than 1,500 Spanish-English bilingual children from kindergarten to third grade, Miller et al. (2006) found that oral language ability in Spanish predicted Spanish reading scores and that oral language ability in English predicted English reading scores. Cross-language effects were also found in that oral language scores in one language

predicted reading scores in the other language. This suggests that oral language skills are highly related to academic skills within and across languages.

Narratives are ideal for evaluating language development since they are able to encourage the variety and complexity of language (e.g., syntax and vocabulary) productions found in the contexts of daily communication. Studies have shown that oral narratives generate longer utterances as compared to conversation (Leadholm & Miller, 1992; MacLachlan & Chapman, 1988) and expository speech (MachLachan & Chapman, 1988). Moreover, possibly due to its higher level of communicative demands, the use of narratives allows for accurate differentiation between typically-developing monolingual children and those with language delays (MachLachan & Chapman, 1988). MacLachlan and Chapman (1988) examined the communication breakdowns (including “stalls,” “repairs,” and “abandoned utterances”) (Hieke, 1981) in the speech of seven children with language disabilities. They found that these children exhibited significantly more communication breakdowns per communication unit in a narrative situation than in conversation as compared to controls.

Studies also have demonstrated that certain skills in the context of narratives of children with language delays are significantly lower than those of typically developing children (e.g., Boudreau & Hedberg, 1999; Kaderavek & Sulzby, 2000; Scott & Windsor, 2000). These skills include total number of words (Boudreau & Hedberg, 1999; Scott & Windsor, 2000), words per minute (Scott & Windsor, 2000) and morphology (Kaderavek & Sulzby, 2000). For example, Kaderavek and Sulzby (2000) studied the narratives and story book readings of 20 two- to four-year-old preschool children (10 with SLI and 10 typically developing matched on age and

gender). Among their findings, Kaderavek and Sulzby (2000) found that children with language impairment had more difficulty with certain linguistic skills in the context of narratives (e.g., use of past tense and personal pronouns) as compared to typically developing children. The evidence thus far, supports the use of narratives for language assessment.

The oral narrative, in particular, has been found to be quite useful in eliciting language that can be compared across languages, since children are able to talk about the same subject and story sequence (Berman & Slobin, 1994; Pearson, 2002). Berman and Slobin (1994) effectively used the procedure of using wordless picture books in their venerable research on cross-linguistic language development. In this task, the child looks at the picture sequences and narrates the story. Miller et al. (2006) further expand the task by providing the child a model and asking the child to retell the story while looking at the picture book. Miller et al. (2006) argue that ELLs benefit from the narrative retell procedure because many of the children are often not exposed to this type of discourse in the home environment. The presence of a model for producing a story is instrumental since children construct better narratives in a retell situation than when they independently create a story (Liles, 1993). The retell context is one of high speaking demand in that it requires components such as specific information, good sentence planning, and accurate lexical retrieval in order to provide logical and organized information.

Via these narratives, linguistic skills can be generally evaluated. For example, level of syntactic complexity can be determined by calculating mean length of utterance (MLU) (e.g., Brown, 1973). Vocabulary skills can be ascertained by

measuring lexical diversity of their utterances (e.g., number of different words/NDW) (e.g., Miller, 1991; Klee, 1992; Klee, Stokes, Wong, Fletcher, & Gavin, 2004).

Additionally, overall language fluency can be assessed by calculating number of words per minute (Riggenbach, 1991). Researchers assert that oral language measures obtained using the analysis of language sample, such as mean length of utterance (MLU), Number of Different Words (NDW) and words per minute (WPM) are a less-biased means of language evaluation for ELLs (e.g., Miller et al., 2006) as compared to traditional standardized assessments.

The majority of studies on the developmental progression of these oral language measures have involved only English-speaking children (e.g., Rice, Wexler, & Hershberger, 1998; Tilstra & McMaster, 2007). Those studies done in other languages have also focused on monolingual individuals (e.g., Klee et al., 2004). The few studies examining the productive language measures in Spanish-speaking ELLs usually involve cross-sectional data (Miller et al., 2006; Muñoz et al., 2003), small sample sizes, and evaluation primarily of only one language (e.g., Muñoz et al., 2003). Use of alternative evaluation procedures such as narrative language sampling is a major historical step in the development of appropriate assessment of Spanish-speaking ELLs. However, this effort must be furthered by determining appropriate methods of norming of measures derived from the narrative. The following section will identify the measures derived from narratives that the literature on oral language developmental recommends for charting language development. These measures will be used to address issues related to norming.

In the process of deciding upon the narrative measures most ideal for looking at language development, researchers have created two broad categories of narrative structure called “macrostructure” and “microstructure” (Hughes et al., 1997; Owens, 1999; Paul, 2001). Although there are no strict guidelines for determining the most useful outcome variables for narrative study, it is recommended that elements of both “macrostructure” and “microstructure” be evaluated (e.g., Hughes et al., 1997; Owens, 1999; Paul, 2001). Macrostructure pertains to overall narrative structure (e.g., story grammar), while microstructure concerns linguistic elements (e.g., syntactic structure). While both macrostructural and microstructural properties of the narratives are important, studies have shown that microstructural measures are sensitive in gauging the levels of children’s linguistic skills as well as in identifying those with language delays (e.g., Scott & Windsor, 2000; Van der Lely, 1997). For example, Scott and Windsor (2000) examined the ability of 10 general language performance measures to differentiate between school-age children with language learning disabilities and their chronological age and language age peers. Children with language learning disabilities who provided oral and written summaries of two educational videotapes scored significantly lower on measures including t –units (a T-unit is a dominant clause and its accompanying independent clauses; Hunt, 1965), total words, and words per minute. Fluency (percent T-units with mazes) and lexical diversity (number of different words) were comparable for all the children. This information demonstrates that all general performance measures might not be sensitive to language development or of differentiating between typically developing children and those with language disorders.

Related to microstructure, general outcome indicators (GOIs) (Deno, 2003; Deno, Mirkin, & Chaing, 1982; Fuchs, 2004; Fuchs, Fuchs, & Speece, 2002) are quick measurements that are used repeatedly over time to gauge progress using a standard method of elicitation. Results of GOIs are highly related to performance in a larger area of functioning (e.g., reading) (see Wayman, Wallace, Ticha, & Espin, 2007 for a review). In the area of language development, Tilstra and McMaster (2007) found that two measures (“total productive words per minute” and “total number of words per minute”) were reliable in determining verbal fluency. In their study of 250 five- to 12-year-old children using a single picture elicitation task for the development of the Index of Narrative Microstructure, Justice, Bowles, Kaderavek, Ukrainetz, Eisenberg, and Gillam (2006) identified language output variables that represented two factors (Productivity and Complexity). The Productivity factors included word output, lexical diversity, and T-unit output while the Complexity factor dealt with syntactic organization, which included mean length of T-units in words and proportion of complex T-units. Although the factors were related, each pointed to specific aspects of expressive language skills. This suggests that specific areas of language must be assessed in order to obtain a complete profile of children’s language development.

There are few studies that address indicators of language development in children’s two languages. Miller et al. (2006) investigated language output measures in over 1,500 Spanish-English kindergarten to third grade bilingual children in order to assess the ability of these measures to predict reading ability within and across languages. These measures included mean length of utterance, number of different

words, and words per minute. They stated that these assessment measures have been proposed to be equivalent across English and Spanish and contain less test bias (inability to provide appropriate evaluation results, due to limited sensitivity and specificity) as compared to standardized assessment procedures. As noted previously, results of their study indicated that oral skills in one language predicted reading scores within languages, as well as crosslinguistically.

Complexity, productivity, and fluency are measuring different aspects of language (e.g., MLU measures syntactic skills, NDW measures semantic ability, while WPM measures language fluency). Attesting to their distinct measurement capacities, MLU (morphemes) and NDW were not correlated when age was controlled (Klee et al., 2004). Moreover, the literature suggests that combinations of these oral measures can validly differentiate between children with and without language deficits (e.g., Klee et al., 2004). With regard to ELLs, MLU, NDW and WPM are considered the most robust measures in bilingual language research (e.g. Miller et al., 2006). Moreover, these are the easiest of the GOIs to calculate using programs such as SALT (Miller & Iglesias, 2007)

While such measures as MLU, NDW, and WPM have generally been found to more appropriately characterize the language development of children of the Spanish-speaking ELL population (in contrast to conventional standardized instruments), there have been few studies conducted on the developmental progression of these language measures in ELLs. It is imperative that the trajectories of these language variables in this population be determined because it will help to define a continuum of normalcy

and subsequently used to accurately differentiate between typical and atypical language development.

Mean Length of Utterance

MLU in morphemes has enjoyed a long history for use in charting children's language development (e.g., Bedore & Leonard, 1998; de Villiers & de Villiers, 1973; Miller & Chapman, 1981; Paul & Alforde, 1993; Thordardottir, 1998). Brown (1973) suggested that MLU better characterized the language development of children than age given the highly variable rate of language acquisition. However, other studies have demonstrated that age and MLU are highly correlated (Blake, Quartaro, & Onorati, 1993; Klee, Schaffer, May, Membrino, & Mougey, 1989; Miller & Chapman, 1981). An important feature of MLU is its ability to be used to distinguish between typically developing children and those with language delays or disorders (e.g., Klee et al, 2004). Klee et al. (2004) examined the conversational language skills of Cantonese children across two studies. In their second study with 45 children (15 with Specific Language Impairment (SLI), 15 typically developing children matched on age and another 15 matched on language comprehension), they found that measures of utterance length (MLU) and lexical diversity were able to distinguish between Cantonese children with typical language development and those with SLI. Children with SLI produced significantly shorter utterances (Klee et al., 2004). The research supports MLU as a sensitive indicator of language development.

Most studies on MLU focus on morphemes instead of words. However, it should be noted that MLU in morphemes is highly correlated with MLU in words

(e.g., Thordardottir, 1998) so both can be used as indicators of language development. However, MLU in morphemes is not appropriate for cross- language comparisons due to the differences in language structures which could cause the measure to be inconsistent across languages. For example, since French is a highly inflected language (Comeau, Genesee, & Lapaquette, 2003), MLU in words is the preferred calculation when making comparisons with English. Likewise, due to the highly inflected nature of the Spanish language, it is more accurate to use MLU in words as a measure when making cross-language comparisons with less inflected languages like English (Guasti, 2004).

Number of Different Words

Lexical diversity indices utilizing language samples have been frequently used to assess vocabulary size in children (e.g., Miller, 1991; Klee, 1992; Klee et al, 2004). Measures of lexical diversity include number of different words, which has been found to be lower in children with delayed language skills as compared with their typical language peers matched on age for different languages (e.g., French) (Thordardottir & Namazi, 2007) and English (Watkins, Kelly, Harbers & Hollis, 1995). NDW was also found to significantly correlate with age in a linear manner (Miller, 1991), and was able to predict reading scores in bilingual children (Miller et al., 2006). Thordardottir and Namazi (2007) examined spontaneous language samples of 12 French-speaking children with SLI, and 12 typically language developing children matched on age and 12 matched on MLU in order to examine areas of strengths and weaknesses of children with SLI. They found that children with SLI

performed lower on measures of utterance length, lexical diversity and composition as compared to their age-matched peers. However, they scored comparably to their MLU- matched controls.

Watkins et al. (1995) studied 75 monolingual English-speaking preschool children (25 with SLI, 25 typically developing children with similar language ability as those with SLI, and 25 typically developing children matched to the children with SLI based on chronological age) to compare the sensitivity of type token ratio and number of different words in differentiating children with typical development from those with SLI. They found that children with SLI produced a significantly lower number of different words than their age-matched peers. This was the case both when the samples were matched on utterance length and when they were not (Watkins et al., 1995). This indicates the usefulness of NDW for determining the development of lexical diversity.

Conflicting findings are present concerning NDW, however. In English-speaking children, Scott and Windsor (2000) found that NDW did not differentiate between children with language impairment from their language age and chronological age peers while examining the ability of 10 general language performance measures to differentiate between school-age children with language learning disabilities and their chronological age and language age peers.

Regarding Spanish-speaking children, Muñoz et al. (2003) found that NDW and total number of words were not sensitive to language development for young Spanish-speaking children from low socioeconomic backgrounds. Muñoz et al. (2003) studied measures of language productivity, sentence organization, and story

structure in the narratives of 24 Latino children from low socioeconomic situations (grouped according to 12 younger and 12 older children). They found that the narratives of older children had longer sentences, more grammatically acceptable sentences, and more complete narrative episodes as compared to younger children. However, they found that language productivity measures (including total number of words and number of different words) did not differentiate between the two groups of children. They concluded that these measures are not sensitive markers of language development in this population due to large variability in story length. The lack of significance might be due to the small number of individuals in the study. Overall, discrepancies in the literature concerning NDW might be due to methodological differences among the studies; specifically, sample size.

Words Per Minute

Language productivity measures have been shown to differentiate between children with high and low language ability (Allen & Bliss, 1987; Allen, Kertoy, Sherblom, & Pettit, 1994) and between children with and without language disorders (Crais & Lorch, 1994; Paul & Smith, 1993). Tilstra and McMaster (2007) studied “measures of language productivity,” “verbal fluency,” and “grammaticality” in 45 kindergarten, first-grade, and third-grade children using elicited narratives. They found that the measure “words per minute” was able to differentiate third grade children from those in kindergarten and first grade. Moderate criterion validity was also established between WPM and a standardized oral language measure. Words per minute also discriminated between school age children with and without language

impairment (matched on age) (Scott & Windsor, 2000). Overall, WPM has been established as a good indicator of language development.

These oral language measures (MLU, NDW, and WPM) have been used with Spanish-speaking ELLs. However, existing norms are limited to those collected by Miller and Iglesias (2007) as part of the Bilingual Language and Literacy Project (BLLP). The BLLP project, part of a larger project, includes a very large database of oral narratives, considered to be one of the factors that can contribute to the variability in the literacy development of Spanish-speaking ELLs. Research based on data from this project, indicate that MLU, NDW, and WPM have been demonstrated to be reliable indices of language development in ELLs (Heilmann et al., 2008; Miller et al., 2006). However, as noted previously, current norms involving MLU, NDW, and WPM are based on cross-sectional data. In addition to the issue of which particular language variables to include in the norming of oral narratives of ELLs, an important concern in norming of language measures is whether age or grade norms should be used. This is one of the questions of this project. The following section will examine the use of age and grade norms in education and developmental language literature.

Research Questions

Norming by Age or Grade?

In order to assess language skills of school-age children, the effects of age and education must be considered (Alexander & Martin, 2004). Standardized oral language tests may present norms by age or grade. However, studies in some areas of

child development have reported that using norms based on age creates biases related to grade level. More specifically, there are significant differences in performance scores when age and grade variables are used, with scores using age as the index of time overwhelmingly being lower than scores using grade as the index of time. The implications of the research studies are that exposure to the educational system has a significantly higher effect on children's educational scores (e.g., emergent literacy and reading) than the effects of age (Crone & Whitehurst, 1999; Morrison, Griffith, & Alberts, 1997). Despite the advantage of using grade in the norming of certain tests, it is still customary to norm by age, partially because age is often used as the index of time in most developmental studies. Another reason for the preference for age norms is due to perceived ease of interpretation (Alexander & Martin, 2004). Age and grade are highly correlated (Alexander & Martin, 2004) and are comparable for older children in the area of reading. However, in the early years the effect of one year can be considerable, especially at a time when important decisions are being made regarding academic placement. Additionally, age and grade norms differ in how they deal with children who are outside the standard age range within a grade (children who are kept back a grade or given advanced placement at school entry or at promotion). The concern is that when grade norms are used, there can be substantial variation in children's ages at each grade level. For example, in some studies, there were some children at each grade level that were a year or more apart in age (e.g., Cahhan & Cohen, 1989). Likewise, when age norms are used, children of the same age groups can pertain to different grades. As a result, age and grade norms might not be comparable.

The effects of age, grade, or both, on development have been investigated by a number of researchers, for example in the areas of literacy and mathematics (Morrison et al., 1997) and in the area of verbal cognitive ability (Cahan & Cohen, 1989). In order to study the effects of entrance age on reading and mathematics achievement in first grade children, Morrison et al., examined pre- and post performance of 539 younger first graders, older first graders, and older kindergarteners (for oldest and youngest two-month age groups within age at each grade level). They found that younger first graders made comparable progress to older first graders, and made considerably more progress than older kindergartners. They concluded that entrance age alone was not a good predictor of academic improvement or delay.

Research in the area of verbal cognitive ability with monolingual children (Cahan & Cohen, 1989) found that the effect of one grade was more than twice the effect of one year of age for the majority of verbal cognitive ability subtests in fourth to sixth grade children in Hebrew language schools in Jerusalem. These findings were based on separating the age from grade effect, because age and grade are highly correlated. Findings in the area of reading also report similar effects, with some researchers highlighting the importance of separating the effects of age and grade and studying the age within grade effect (e.g., Alexander & Martin, 2004; Cahan & Cohen, 1989; Cahan & Noyman, 2001; Crone & Whitehurst, 1999).

Crone and Whitehurst (1999) studied the effects of age and schooling on the emergent literacy and early reading skills of 337 children from low-income backgrounds from the end of Head Start to the end of 1st grade. One hundred eighty

three of these children were also followed until the end of second grade. They found that children who began school earlier than peers of the same age, performed better in emergent literacy and reading skills than their peers. Results of the study indicated that the effect of a year of schooling on literacy skills was 1.7 times greater than development related to age. Additionally, the effect of a year of school on early reading was 4.3 times greater than that of age. Crone and Whitehurst (1999) used a “cutoff design” in which they included only children who adhered to the Suffolk County, New York guidelines for kindergarten enrollment which specified that children must be five years old prior to December 1st of their enrollment year. They excluded individuals that were born after the cutoff date but began school early, as well as children who were born before the cutoff, but began school late. The researchers utilized the 12-month age range within a grade level to make comparisons between two groups of children with the same length of exposure to an academic environment and who have different age levels. They suggested that these comparisons can separate the effect of an added year of schooling on academic skills. Further analyses were made using a “regression discontinuity design” (Crone & Whitehurst, 1999) which allowed for the evaluation of the effects of age on the different measures. The regression slopes indicated the effects of age, while the discontinuity between two regression lines indicated the effect of schooling.

Alexander and Martin (2004) studied the effects of age and grade for the reading mastery of 4,257 first and second grade children in a school that adheres to a stringent policy of age to grade assignment (in Tasmania). Children already placed within their strict age for grade limits were divided into four three-month

subcategories based on age of enrollment. Studying the age within grade effect, using three subtests of the Woodcock Reading Mastery Test, they found significance for both age and grade. However, the effect of grade was approximately twice that of age. The effect of grade and age are expected to decrease with later grades (Alexander & Martin, 2004). Using age-based norms for reading and other verbal tests might give biased results during early school years (Alexander & Martin, 2004). Age-based norms might obscure the true achievement of children because they reflect lower scores as compared to grade-based norms. Therefore, once children enroll in the educational system, developmental variables should be charted by grade because aspects of development typically studied by age “have less influence on children’s verbal performance” (Alexander & Martin, 2004).

Many educational tests use age and/or grade norms. Some tests that include both sets of norms are the Norris Educational Achievement Test (NEAT) (Switzer & Gruber, 1992), The Woodcock Johnson Psycho-Educational Battery-revised (WJ-R) (Woodcock & Johnson, 1989), the Kaufmann Educational Achievement Tests-Revised (KTEA-II) (Kaufman & Kaufman, 2004), the Wechsler Individual Achievement Test – Second Edition (WIAT-II) (Weschler, 2001), and The Gray Oral Reading Test, Fourth Edition (GORT-4) (Weiderhold & Bryant, 2001).

Age norms are widely used in standardized oral language tests (e.g., CELF-4 English & Spanish, EOWPVT (English & bilingual editions)). The Peabody Picture Vocabulary Test IV (Dunn & Dunn, 2007) is one exception, providing separate age and grade norms. Since norms for age and grade differ in how they deal with children who fall outside the typical age for grade (includes children who are given advanced

placement or held back at the beginning of school or as they progress through the grades), both types of norms should be included in normative studies. It is important to realize that age and grade concerns could be exacerbated when dealing with ELLs, who might have a high propensity for entering school at varying ages or of being retained (Heubert & Hauser, 1999; Uriate, Lavan, Agusti, & Karp, 2009). It is important then to determine the contribution of age and grade to development of language variables in the Spanish-speaking ELL population. Age and grade will be used as independent variables for this investigation. Another important question regarding criteria for creating norms is whether cross-sectional or longitudinal data should be used. The next section will address this issue.

Cross-sectional or Longitudinal Data?

Many researchers point out that using cross-sectional instead of longitudinal data is disadvantageous. For example, there is a lack of statistical control and power (e.g., Collier, 1992; Hedeker & Gibbons, 2006). However, substantial difficulty can be encountered with longitudinal studies (e.g., Collier, 1992; Hedeker, 2006), such as problems of missing data due to participant attrition.

Cross-sectional studies can be detrimental when caution is not utilized in using them for making inferences about longitudinal development. In the area of child language, in particular, there have been concerns regarding the comparability of cross-sectional and longitudinal research (e.g., Rosansky, 1976). Despite the complexity encountered with collecting and analyzing longitudinal data, however,

there is overall agreement that it is a more accurate method for charting developmental information (e.g., Collier, 1992; Hedeker & Gibbons, 2006).

Kraemer, Yesavage, Taylor, and Kupfer (2000) emphasize the need for caution in making inferences based on cross-sectional studies. They point out methodological differences between studies that can exacerbate the differences in the relationship between cross-sectional and longitudinal studies. Kraemer et al. (2000) conducted some simulated studies based on real situations in order to present clear relationships between cross-sectional studies and accurate longitudinal inferences. They concluded that certain questions can be answered by longitudinal studies only, that cross-sectional studies are very useful in studies that do not involve longitudinal inferences (Kraemer et al., 2000), and that great caution must be used when using cross-sectional studies for making inferences regarding longitudinal outcomes.

There are several advantages of longitudinal data. First, these data provide greater statistical power than is possible with cross-sectional data since fewer subjects are required. “The net result is that the repeated measurements from a single subject provide more independent information than a single measurement obtained from a single subject” (Hedeker & Gibbons, 2006, p.1). Secondly, each subject acts as his/her own control. There is less intra-subject variability than inter-subject variability, so tests based on longitudinal data (having only intra-subject variability) yield statistically superior information to that obtained from cross-sectional studies.

Third, studies using longitudinal data provide for the separation of “aging effects (i.e., changes over time within individuals), from cohort effects (i.e., differences between subjects at baseline)” (pp 1-2). Lastly longitudinal analysis is

able to document change within an individual, while cross-sectional analyses are incapable of this. This latter information can provide information about the heterogeneity of the population being studied as well as examine change within individuals. Some difficulties with longitudinal data analysis include problems selecting the appropriate analysis methods, difficulty with management of the complicated analyses, concerns about the possible changes of values of predictors over time (not just the outcome variables, necessitating complicated models).

Despite the superiority of longitudinal data in most situations, there might be a good rationale for using cross-sectional data. The problem of attrition or difficulty maintaining the original sample for longitudinal study can be severe, particularly when dealing with the ELL population due to their high level of mobility (e.g., Reese, Gallimore, & Guthrie, 2005). Therefore, it would be advantageous if comparable results regarding language development can be found using cross-sectional and longitudinal studies. It would facilitate investigations conducted with the ELL population if cross-sectional studies would be as adequate since they are less time consuming. If longitudinal data are used, however, there are other considerations that must be made in order to determine the most effective use of these data. One concern is the presence of missing data and how it might affect study outcomes. Another is whether certain participants should be included or excluded when they present with certain characteristics that are different from those of the majority of the study sample. The following section will address these two concerns.

*Does Inclusion or Exclusion of Children With Missing Data or Grade Repeats Affect
Language Norms?*

The composition of the dataset is critical in the analysis of developmental data. It is strongly recommended in most instances that the full longitudinal dataset be used because listwise deletion can cause the loss of power and may bias results if participants with complete data are not representative of the complete sample (Graham, Hofer & Piccinin, 1994; Little & Rubin, 2002). However, listwise deletion might not negatively impact very large datasets. Using the complete data allows for an array of missing data analyses which are helpful in explaining the mechanisms of missingness (e.g., Fichman & Cummings, 2003; Schafer & Graham, 2002).

Effective statistical techniques for the analysis of longitudinal data with missingness include maximum likelihood (Duncan et al., 2006; Schafer & Graham, 2002) and multiple imputation (Fichman & Cummings, 2003; Schafer & Graham, 2002; Wayman, 2003). Such techniques are important for accurate and efficient results with longitudinal data containing missing items. If the deletion of missing items does not affect the results of longitudinal analyses, then listwise deletion can be adequate. This project will examine different longitudinal datasets to determine whether there are any substantial differences in the language measures as a function of grade.

Summary

There are numerous issues that must be addressed in a study of determining the optimal procedures for developing language norms for ELLs. There has been

increasing and considerable concern that in light of the growing population of ELLs there should be appropriate language evaluation instruments to address their needs. Traditionally, standardized tests have not included adequate representation from culturally and linguistically diverse children in their norming samples (Laing & Kamhi, 2003); this is especially true for Spanish-speaking ELLs. Additionally, the range of language proficiency of ELLs has not been adequately addressed with the norming of these instruments.

The literature has established that even well-developed standardized tests cannot, on their own, sufficiently and adequately elicit the language skills of ELLs. Therefore, it has been widely suggested that alternatives, in particular, oral narratives, be used to examine their language development (e.g., Miller et al., 2006). However, norms for narrative measures used with ELLs are lacking. Substantial progress has been made towards the developing of these norms, such as agreement on specific measures derived from narratives that should be used. The most effective measures include MLU, NDW and WPM (e.g., Heilmann et al., 2008; Miller et al., 2006). However, norms available for these measures are based on small scale studies (e.g., Muñoz et al., 2003) and cross-sectional data (Muñoz et al., 2003; Miller et al., 2006). The Miller et al. study is based on a very large cross-sectional database of ELLs.

The addition of longitudinal data would provide much needed information on the development of these language variables in ELLs. However, with the presence of longitudinal data there are several remaining issues. These involve whether age or grade should be used as an index of time with language variables, whether cross-sectional or longitudinal data should be used in the development of norms for MLU,

NDW, and WPM, and whether including or excluding children with missing data or who have repeated a grade, affects language norms.

Regarding the first of these issues, the literature has shown that the grade variable is more appropriate for evaluating certain areas of child development (for example, literacy development), especially in the early school years (e.g., Alexander & Martin, 2004). However, this type of information is not available for the oral language variables, particularly in ELLs. Concerning the second issue, longitudinal data has been recommended over cross-sectional data for examining aspects of development (e.g., Collier, 1992; Hedeker & Gibbons, 2006). However, with the possibly high attrition rate of ELLs during longitudinal studies, it would be advantageous to determine if there are substantial differences between results of cross-sectional and longitudinal analyses for the language variables being studied. As for the third issue, regarding whether including or excluding children with missing data, or who have repeated a grade, affects language norms for ELLs, no such information is available. However, research in other areas of recommend detailed analyses of missing data, as well as the inclusion of cases with missing data in longitudinal analyses (e.g., Fichman & Cummings, 2003; Schafer & Graham, 2002).

This research project is concerned with determining appropriate methods for establishing norms for Spanish-speaking ELLs. From the review of the literature, there are several issues that must be addressed. These issues are whether age or grade norms should be utilized, whether cross-sectional or longitudinal data should be used, or whether the inclusion or exclusion of children with missing data or grade repeats affects the language measures. Due to the syntactic and lexical differences that exist

among languages, an identical developmental schedule is not expected between the English and Spanish of ELLs.

The questions of this research project are the following:

- Should age or grade norms be used as an index of time to examine the development of MLU, NDW and WPM? It is hypothesized that grade results will provide a higher level of significant differences than age results.
- Should cross-sectional or longitudinal data be used in the development of norms for MLU, NDW, and WPM? It is hypothesized that results of longitudinal data will provide a higher level of significance than those of cross-sectional data.
- Does inclusion or exclusion of children with missing data or grade repeats affect language norms? It is hypothesized that there will be no significant differences between datasets including or excluding children with missing data or grade repeats, especially due to the large datasets that are being utilized.

CHAPTER 2

METHODOLOGY

The overall aim of the present study is to investigate appropriate norming procedures for determining the developmental progression of English and Spanish language skills of English Language Learners. A series of research questions have been posed to address this aim. The first question to be addressed is whether age or grade variables should be used as an index of time for evaluating MLU, NDW and WPM. The second question deals with the issue of whether norms vary as a function of using cross-sectional or longitudinal data. The third question examines whether including or excluding children with missing data or those who have repeated a grade affects the language norms. To answer the proposed questions, existing cross-sectional and longitudinal data from typically developing ELLs enrolled in transitional bilingual programs were used.

Participants

Participants are Spanish-speaking English Language Learners (ELLs) enrolled in kindergarten to second grade in Texas. The participants are part of a larger project, referred to as the Bilingual Language and Literacy Project (BLLP), investigating factors that contribute to the variability in reading and school achievement in Spanish-speaking ELL children (Francis, Carlson, Fletcher, Foorman, Goldenberg, & Vaughn et al., 2005). The BLLP database consists of two distinct projects. The first project is a cross-sectional study of kindergarten to third grade ELLs enrolled in

transitional bilingual programs. The second project is a longitudinal study of kindergarten to second grade ELL children enrolled in a variety of bilingual programs in Texas and California who were assessed twice a year for three years. In order to control for school program, only children enrolled in transitional programs were selected. Additionally, only children who participated at least in one of the three data collection points (the fall of kindergarten, first grade, and second grade) were included. The children in both projects attended programs in two geographic areas of Texas (urban south-eastern Texas and the Rio Grande Valley- border between Texas and Mexico). An inclusionary criterion for the BLLP project was that the children be typically developing as indicated by non-enrollment in special education programs.

At the inception of the data collection portion of the projects, English language skills of children in both projects were judged by their school district to be inadequate to be able to function well in English only classrooms. The transitional program in which the children were enrolled consisted of primary instruction in Spanish, with gradual transitioning to English. There was some variation in the transition programs among the schools; some used an early transitional model while others followed a late transition model. Early transition programs moved children to English only classrooms after third grade, while late transition programs made the change after fourth grade.

Cross-sectional Sample

The original BLLP database consisted of 928 children. Only children enrolled in kindergarten, first, and second grade who were able to produce at least one four-

utterance narrative sample in either English or Spanish, participated in this study. Information was unavailable regarding whether children repeated a grade. All 299 third graders were eliminated from the original database. Seven children whose ages appeared to have been incorrectly entered in the database were also deleted (two years older than the oldest child in the final grade sample for a particular grade level). Two distinct datasets (cross-sectional grade dataset and cross-sectional age dataset) were constructed for each language. The cross-sectional grade dataset consisted of all children enrolled in kindergarten first, and second grade, regardless of their age. The 605 children in the cross-sectional grade dataset consisted of 138 children in kindergarten, 215 in first, and 252 in second grade. There were 282 males (47 percent) and 323 females (53 percent). Mean ages for the kindergarten, first, and second grade children for English samples were 6.0 (SD= .36), 7.0 (SD= .38), and 8.1 (SD = .48), respectively. For Spanish samples, the mean ages for the kindergarten, first, and second grade were 5.9 (SD= .37), 7.0 (SD = .39), and 8.0 (SD = .49), respectively. The range in age for kindergartners, first, and second grade in English were 4.2-6.7, 6.3-8.2, and 6.4-9.9, respectively. The age range for kindergartners, first grade and second grade in Spanish were 4.1 to 6.6, 6.1 to 8.2, and 6.3 to 9.9, respectively.

The cross-sectional age dataset consisted of children in each grade level that were six months above or below the mean age for the particular grade. Children outside of the age range for their particular grade were deleted (15 children or 11 percent of children enrolled in kindergarten, 36 children or 17 percent of children enrolled in first grade, and 49 children or 19 percent of children enrolled in second

grade. As expected, the largest group of children deleted because of being out of age range involved those enrolled in second grade). The final sample of 505 participants consisted of 123 children in kindergarten, 179 in first, and 203 in second grade. For the English sample, the mean age of kindergartners within the age range of 5.5 to 6.6 was 6.1 (S.D = .27), the mean age of first graders within the age range of 6.5 to 7.6 was 7.1 (SD = .30) and the mean age of second graders within the age range of 7.5 to 8.6 was 8.1 (SD=.31). For the Spanish sample, the mean age of kindergartners within the age range of 5.5 to 6.5 was 6.0 (SD = .28), the mean age of first graders within the age range of 6.5 to 7.5 was 7.0 (SD = .29), and the mean age of second graders within the age range of 7.5 to 8.5 was 8.0 (SD=.30).

Longitudinal Datasets

Three longitudinal datasets were constructed from the full BLLP longitudinal database. The 679 children selected were kindergarten to second grade children enrolled in transitional bilingual programs. The first dataset, referred to as Longitudinal Dataset I consisted of all the children, including those who had data missing at one or more of the three data collection points and who repeated a grade. The second dataset, Longitudinal Dataset II, consisted of all the children except those who repeated a grade. The third dataset, Longitudinal Dataset III, consisted of children who had data at all three data collection points and who did not repeat a grade. What follows is a description of the children in each of the three datasets.

Longitudinal Dataset I: All 679 children enrolled in transitional programs in kindergarten, first grade, and second grade constituted the Longitudinal Dataset I. Longitudinal Dataset I included children who did not participate at all collection points.

Missing data summaries for English and Spanish can be found in Appendix A. These provide information regarding the number and frequency of patterns involved in the missing data. Gender data were available for 635 of the 679 children. Of the available data, there were 321 Males (51%) and 314 females (49%). As can be seen in Table 3 there are differences in the number of children at each grade level and in the number of children who produced narratives in English and Spanish. The number of children who were able to produce narratives in English and Spanish increased as a function of grade. As expected due to their ELL status, the number of children who produced narratives only in Spanish was larger than the number of children who produced narratives in English, especially in kindergarten. The mean ages for kindergartners, first and second graders in English were 5.7 (SD = .32), 6.7 (SD = .37), and 7.7 (SD = .39), respectively. The mean ages for kindergartners, first and second graders in Spanish were 5.6 (SD = .33), 6.7 (SD = .37), and 7.7 (SD=.39), respectively. The age ranges for kindergartners, first, and second grade in English were 5.1 to 6.6, 6.1 to 8.3, and 6.8 to 9.3, respectively. The age ranges for kindergartners, first, and second grade in Spanish were 5.1 to 7.1, 5.8 to 8.3, and 6.8 to 9.3, respectively.

Table 3. Number and percentage of children who produced narratives in both English and Spanish, English only, and Spanish only at each data collection point (Longitudinal Dataset I)

| Grade | Number of Children in Grade | English & Spanish | English Only | Spanish Only |
|-------|-----------------------------|-------------------|--------------|--------------|
| K | 270 | 121 (45%) | 14 (5%) | 135 (50%) |
| 1 | 457 | 363 (79%) | 22 (5 %) | 72 (16%) |
| 2 | 470 | 441 (94 %) | 10 (2%) | 19 (4 %) |

Longitudinal Dataset II: This dataset consisted of all children in the original longitudinal dataset (Longitudinal Dataset I) who did not repeat kindergarten or first grade (N= 639). A total of 40 children had repeated a grade (nine in kindergarten and 31 in first grade). Gender information was present for 595 children. There were 297 (50%) males and 298 (50%) females. Table 4 demonstrates the percentage of the children who produced both English and Spanish narratives, only English narratives, only Spanish narratives.

The mean ages for kindergartners, first and second graders in English were 5.6 (SD = .32), 6.7 (SD = .37), and 7.7 (SD = .39), respectively. The mean ages for kindergartners, first and second graders in Spanish were 5.6 (SD = .32), 6.7 (SD = .37), and 7.7 (SD=.39), respectively. The age ranges for kindergartners, first, and second grade in English were 5.1 to 6.6, 6.1 to 8.3, and 6.8 to 9.3, respectively. The age ranges for kindergartners, first, and second grade in Spanish were 5.1 to 7.1, 5.8 to 8.3, and 6.8 to 9.3, respectively.

Table 4. Number and percentage of children who produced narratives in both English and Spanish, English only, and Spanish only at each data collection point (Longitudinal Dataset II)

| Grade | Number of Children in Grade | English & Spanish | English Only | Spanish Only |
|-------|-----------------------------|-------------------|--------------|--------------|
| K | 251 | 112 (45%) | 14 (6 %) | 125 (49%) |
| 1 | 427 | 339 (79 %) | 20 (5%) | 68 (16%) |
| 2 | 437 | 411 (94%) | 10 (2%) | 16 (4 %) |

Longitudinal Dataset III: The third dataset consisted of children from the original sample of 679 children (Longitudinal Dataset III) who had not repeated a grade and for which data were available at each of the three data collection points. There were 74 children with English samples and 163 children with Spanish samples at each of the three collection points. For the English group, there were 25 males (34%) and 49 females (66%) and for the Spanish group there were 74 males (44%) and 89 females (55%). The mean ages for kindergartners, first and second graders in English were 5.7 (SD = .30), 6.7 (SD = .30), and 7.7 (SD = .30), respectively. The mean ages for kindergartners, first and second graders in Spanish were 5.6 (SD = .30), 6.7 (SD = .30), and 7.7 (SD=.30), respectively. The age ranges for kindergartners, first, and second grade in English were 5.2 to 6.5, 6.1 to 7.6, and 7.1 to 8.6, respectively. The age ranges for kindergartners, first, and second grade in Spanish were 5.1 to 6.6, 6.1 to 7.6, and 7.1 to 8.6, respectively.

In order to develop comparable datasets that could be used in the comparison of cross-sectional versus longitudinal analyses, two distinct longitudinal datasets (longitudinal grade dataset and longitudinal age dataset) in each language were

created for Longitudinal Dataset III. The procedure used to create the two datasets was identical to the one used to create the age and grade datasets of the cross-sectional dataset. The Longitudinal III grade dataset consisted of all children enrolled in a particular grade, regardless of age. The Longitudinal III age dataset consisted of children in each grade level that were six months above or below the mean age for the particular grade. For the English sample, the mean age of kindergartners (N= 72) within the age range of 5.2 to 6.2 was 5.7 (S.D = .28), the mean age of first graders (N= 72) within the age range of 6.2 to 7.2 was 6.7 (SD = .27) and the mean age of second graders (N=72) within the age range of 7.2 to 8.2 was 7.7 (SD=.28). For the Spanish sample, the mean age of kindergartners (N=157) within the age range of 5.1 to 6.1 was 5.6 (SD = .28), the mean age of first graders (N=157) within the age range of 6.2 to 7.2 was 6.7 (SD = .28), and the mean age of second graders (N= 157) within the age range of 7.2 to 8.1 was 7.7 (SD=.28).

Procedure

Wordless picture books by Mercer Myer were used to elicit the narratives. The examiner was seated across from the child to encourage use of language, and to avoid discrete labeling and the use of pointing. The examiner told a set script of the story in Spanish as the child looked on (See Appendix B for scripts used). Afterwards, the child was asked to retell the story using the wordless, picture book. The examiner remained silent except for backchannelling (e.g., ‘Aha,’ ‘Si’ ‘Tell me more’) (Miller et al, 2006, p. 33) or restated the child’s utterance immediately after it was said in order to encourage the child to continue narrating the story. The

procedure was repeated for the English narrative about a week later. Testing was first done in Spanish since it was presumed to be the children’s stronger language, thereby allowing them to become familiar with the task for a favorable response to the elicitation of the English narrative.

For the cross-sectional sample, the book “Frog, Where Are You?” was used to elicit all the language samples. For the longitudinal sample, various Frog-story books were selected for each collection point, the fall of each school year. The schedule of the books used is displayed in Table 5.

Table 5. Schedule of books for data collection

| Year | Grade | Book |
|-----------|--------------|----------------------|
| 2002-2003 | Kindergarten | Frog, Where Are You? |
| 2003-2004 | First Grade | Frog on His Own |
| 2004-2005 | Second Grade | Frog, Where Are You? |

Transcription and Coding

Digital recordings were made during the narrative collection process. Later the narratives were transcribed by trained assistants using the SALT conventions for transcribing and coding bilingual samples (Miller, & Iglesias, 2007). The SALT conventions were used to accommodate Spanish and Spanish-influenced English. The training process consisted of direct training by the lab manager for approximately 10 hours (Miller et al, 2006). Utterance segmentation was done using C-units (Loban, 1976). A C-Unit was comprised of a main clause and its subordinate clauses. A

deviation from Loban's method was made in the case of coordinated sentences with "ellipted subjects in the second main clause ("la rana brincó y buscó al niño/the frog jumped and looked for the boy") (Miller et al, 2006, p. 34), regarding them as individual C-Units. Loban kept the coordinated units intact since separating would make the sentence semantically absurd. However, since in morphologically rich languages such as Spanish, information regarding the subject is present in the verb, structure and meaning of the utterance is retained. Additionally, young children use "and" frequently (Hunt, 1965) in their sentence production. In Spanish, the frequent use of "and" combined with the high occurrence of ellipted utterances, would cause awkward sentences. This could also cause incorrectly high sentence complexity scores, since the use of multiple compound phrases in a sentence does not truly increase sentence complexity. In order to provide the most accurate measure of sentence complexity, it was decided that coreferential sentences with deleted subject in the second clause would be regarded as a separate C-Unit. Using this modified segmentation method resulted in lower MLU scores than those that are reported in the literature (Rojas, Pereira, & Iglesias, 2000). The same segmentation method was used with Spanish and English samples to make it possible to compare them adequately.

For both the English and Spanish samples, after transcription by one transcriber, another transcriber reviewed the transcriptions and coded the transcripts. The transcripts were analyzed using SALT's rectangular data file procedure, providing MLU, NDW, and WPM scores for each child.

Reliability

Reliability measures were calculated (Miller et al., 2006) at three levels using 40 transcripts (20 in Spanish and 20 in English). “Protocol accuracy” which showed how well transcribers adhered to the rules of SALT (as reviewed by a senior lab manager) “ranged from 98 to 100 percent in English and 94 to 99 percent in Spanish” (Miller et al, 2006, p. 34). “Transcription accuracy” concerned the agreement by two teams of research assistants regarding the transcription accuracy of words, morphemes and “utterance segmentation.” Agreement ranged from “90 to 98 percent in English and 91 to 99 percent in Spanish.” Finally, “coding agreement” (p. 34) calculation for the NSS involving “subjective judgment by the transcribers” (p. 34) resulted in Krippendorff’s alphas of .74 for English and .60 in Spanish (Miller et. al., 2006). “Scaled percent agreement measures for total scores for these coded analyses were greater than 90 percent” (Miller et al., 2006, p. 34).

Furthermore, test-retest reliability performed using 241 transcripts of ELL children (Heilmann, Miller, Iglesias, Fabiano-Smith, Nockerts, & Andriacchi, 2008) indicated that significant correlations were found between two time periods for four narrative measures. These findings indicate that it is possible to accurately transcribe oral narratives and that the measures are consistent, demonstrating its possible clinical utility.

Analysis Procedure

A series of descriptive and inferential analyses were conducted in order to answer the proposed questions. The two languages (English and Spanish) were examined separately.

To answer the first question regarding whether age or grade norms should be used as an index of language development, data from the two cross-sectional datasets (age and grade) and the two datasets of the Longitudinal Dataset III (age and grade) were compared. Multivariate Analysis of Variance (MANOVA) procedures were used to ascertain the effects of age and grade on MLU, NDW, and WPM scores for the cross-sectional dataset. Repeated Measures Analysis of Variance procedures were used to determine these effects with the Longitudinal Dataset III. To answer the second question regarding whether cross-sectional or longitudinal data should be used in the development of norms for MLU, NDW, and WPM, results of the MANOVAs and Repeated Measures ANOVAs conducted for the first question, were compared. To answer the third question regarding whether or not there is an effect on the language measures when children with missing data or those who have repeated a grade are included or excluded, MANOVAS (three in English and three in Spanish) were conducted. Analyses were done to evaluate the performance of the language variables across the three longitudinal datasets (Longitudinal Dataset I, Longitudinal Dataset II, and Longitudinal Dataset III) at each grade level (kindergarten, first grade, and second grade). Comparisons were conducted using means and levels of significance (p values, effect size (partial eta squared values)). Effect size was interpreted using Cohen's recommendations (1988) for eta squared as a guide (.01 =

small, .06 = moderate, .14 large effect). Appendix B contains information regarding how eta squared compares to other indices of effect size (e.g., Cohen's d , r^2 and partial eta squared). For simple analyses, eta squared values are comparable to partial eta squared values, but with more complex analyses (e.g., those more involved than a one-way ANOVA), there is noticeable difference (e.g., Becker, 1999, p 3-4). Additionally, studies using within subject designs are expected to have larger effect size values than those with just between-subject designs (Young, 1993).

CHAPTER 3

RESULTS

The objectives of this study were to determine: (a) whether grade or age variables should be used as an index of time in studies examining the development of MLU, NDW, and WPM, (b) whether cross-sectional or longitudinal data should be used in the development of normative data for MLU, NDW, and WPM, and (c) whether including or excluding children with missing data or those who have repeated a grade affects the language norms. To address the first two objectives, the cross-sectional dataset and the Longitudinal Dataset III, consisting of children who had data at all points and who did not repeat a grade, were analyzed separately for each language. Based on the results of the first set of analyses, the Longitudinal Dataset I (all participants including those who had data missing at one or more points and who repeated a grade), Longitudinal Dataset II (all participants except for those who repeated a grade), and Longitudinal Dataset III were compared in order to address the third objective.

Cross-sectional Dataset

In order to examine whether the means were significantly different across age and grade, MANOVAs and a series of univariate ANOVAS were conducted. The MANOVAs were conducted in order to ascertain whether there were significant differences across age and grade. These were followed by a series of univariate ANOVAs in order to determine for which language variables (MLU, NDW, and WPM) there was significance. Finally, post hoc analyses, using Tukey's

specifications were conducted to determine for which ages and grades there were significant differences. Interpretation of effect size was done using Cohen's guidelines (1988) (.01 = small, .06 = moderate, .14 large effect) for eta squared and Pallant's (2007) recommendation to use the eta squared guidelines as a general guide for partial eta squared. See Appendix C for eta squared and partial eta squared comparisons. Each language was examined separately. The results of the analyses of the English cross-sectional data are presented first, followed by the analyses of the Spanish cross-sectional data.

Examination of Tables 6 and 7 suggests that mean English MLU, NDW and WPM (EMLU, ENDW, and EWPM) tended to increase as a function of age and grade. In order to examine whether the increases were significantly different across age and grade, two MANOVAs and a series of univariate ANOVAS were conducted. The results of the MANOVAs, using Wilk's Lambda, indicated that there was significance for Age ($F(6, 1000) = 11.3, p < .001, \eta_p^2 = .06$) and Grade ($F(6, 1200) = 11.7, p < .001, \eta_p^2 = .06$). Partial eta squared values suggested that effect size was medium for age and grade. In order to ascertain if the effects of age and grade were consistent across the language variables, a series of univariate ANOVAS was conducted. Results of the ANOVAS (Tables 8 and 9) indicated that age and grade were significant for all of the variables.

Age was significant for EMLU, $p = .000, \eta_p^2 = .07$, ENDW, $p = .000, \eta_p^2 = .10$, and EWPM, $p = .000, \eta_p^2 = .05$. Partial eta squared values suggested that effect size was medium for MLU and NDW and small for WPM. In order to further examine the differences, Tukey post hoc analyses were performed. The post hoc analyses

indicated significant differences among some age groups, but not all. As can be seen in Table 10, for EMLU and ENDW there was significance between all of the groups, $p < .05$. For EWPM, the older age group had a significantly higher mean than the other two age groups (5.45 to 6.45 and 6.48 to 7.48). There was no significant difference between age groups 5.45 to 6.45 and 6.48 to 7.48.

Grade was significant for EMLU, $p = .000$, $\eta_p^2 = .06$, ENDW, $p = .000$, $\eta_p^2 = .08$, and EWPM, $p = .000$, $\eta_p^2 = .05$. Partial eta squared values suggested that effect size was medium for MLU and NDW and small for WPM. In order to further examine the differences, Tukey post hoc analyses were performed. The post hoc analyses indicated significant differences among some grades, but not all. As can be seen in Table 11, the post hoc analyses demonstrated that for ENDW, there were significant differences between all of the grades, with higher grades displaying higher scores. For EMLU and EWPM, the means for second grade were higher than those for kindergarten and first grade. No other comparisons were significant.

Table 6. Mean Scores of EMLU, ENDW, and EWPM as a Function of Age (Cross-sectional Dataset)

| Variable | Age group | M | SD | N |
|----------|--------------|-------|-------|-----|
| EMLUW | 5.45 to 6.45 | 6.05 | 1.12 | 123 |
| | 6.48 to 7.48 | 6.39 | 1.0 | 179 |
| | 7.52 to 8.52 | 6.73 | .95 | 203 |
| ENDW | 5.45 to 6.45 | 64.78 | 23.7 | 123 |
| | 6.48 to 7.48 | 76.14 | 24.96 | 179 |
| | 7.52 to 8.52 | 85.44 | 25.10 | 203 |
| EWPM | 5.45 to 6.45 | 70.37 | 24.75 | 123 |
| | 6.48 to 7.48 | 70.85 | 23.64 | 179 |
| | 7.52 to 8.52 | 81.93 | 27.02 | 203 |

Table 7. Mean Scores of EMLU, ENDW, and EWPM as a Function of Grade
(Cross- sectional Dataset)

| Variable | Grade | M | SD | N |
|-----------------|--------------|----------|-----------|----------|
| EMLUW | K | 6.04 | 1.12 | 138 |
| | 1 | 6.29 | 1.01 | 215 |
| | 2 | 6.65 | .99 | 252 |
| ENDW | K | 64.03 | 24.22 | 138 |
| | 1 | 72.88 | 25.17 | 215 |
| | 2 | 82.92 | 26.26 | 252 |
| EWPM | K | 69.51 | 24.62 | 138 |
| | 1 | 68.24 | 23.88 | 215 |
| | 2 | 80.54 | 27.45 | 252 |

Table 8. Univariate ANOVA Results for EMLU, ENDW, and EWPM as a Function of Age (Cross- Sectional Dataset)

| Source | Dependent Variable | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|-----------------|--------------------|-------------------------|-----|-------------|-----------|------|---------------------|
| Corrected Model | EMLUW | 35.715(a) | 2 | 17.858 | 17.436 | .000 | .065 |
| | ENDW | 32947.228(b) | 2 | 16473.614 | 26.980 | .000 | .097 |
| | EWPM | 15450.541(c) | 2 | 7725.270 | 12.056 | .000 | .046 |
| Intercept | EMLUW | 19708.251 | 1 | 19708.251 | 19242.770 | .000 | .975 |
| | ENDW | 2748420.565 | 1 | 2748420.565 | 4501.336 | .000 | .900 |
| | EWPM | 2671112.177 | 1 | 2671112.177 | 4168.572 | .000 | .893 |
| Age | EMLUW | 35.715 | 2 | 17.858 | 17.436 | .000 | .065 |
| | ENDW | 32947.228 | 2 | 16473.614 | 26.980 | .000 | .097 |
| | EWPM | 15450.541 | 2 | 7725.270 | 12.056 | .000 | .046 |
| Error | EMLUW | 514.143 | 502 | 1.024 | | | |
| | ENDW | 306510.562 | 502 | 610.579 | | | |
| | EWPM | 321668.469 | 502 | 640.774 | | | |
| Total | EMLUW | 21511.161 | 505 | | | | |
| | ENDW | 3342233.000 | 505 | | | | |
| | EWPM | 3191993.829 | 505 | | | | |
| Corrected Total | EMLUW | 549.859 | 504 | | | | |
| | ENDW | 339457.790 | 504 | | | | |
| | EWPM | 337119.010 | 504 | | | | |

a R Squared = .065 (Adjusted R Squared = .061)

b R Squared = .097 (Adjusted R Squared = .093)

c R Squared = .046 (Adjusted R Squared = .042)

Table 9. Univariate ANOVA Results for EMLU, ENDW, and EWPM as a Function of Grade (Cross- Sectional Dataset)

| Source | Dependent Variable | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|-----------------|--------------------|-------------------------|-----|-------------|-----------|------|---------------------|
| Corrected Model | EMLUW | 36.416(b) | 2 | 18.208 | 17.380 | .000 | .055 |
| | ENDW | 33381.405(c) | 2 | 16690.703 | 25.834 | .000 | .079 |
| | EWPM | 20635.966(d) | 2 | 10317.983 | 15.755 | .000 | .050 |
| Intercept | EMLUW | 22726.043 | 1 | 22726.043 | 21692.459 | .000 | .973 |
| | ENDW | 3045967.903 | 1 | 3045967.903 | 4714.485 | .000 | .887 |
| | EWPM | 3003460.906 | 1 | 3003460.906 | 4586.082 | .000 | .884 |
| Grade | EMLUW | 36.416 | 2 | 18.208 | 17.380 | .000 | .055 |
| | ENDW | 33381.405 | 2 | 16690.703 | 25.834 | .000 | .079 |
| | EWPM | 20635.966 | 2 | 10317.983 | 15.755 | .000 | .050 |
| Error | EMLUW | 630.684 | 602 | 1.048 | | | |
| | ENDW | 388944.390 | 602 | 646.087 | | | |
| | EWPM | 394254.459 | 602 | 654.908 | | | |
| Total | EMLUW | 25340.369 | 605 | | | | |
| | ENDW | 3829502.000 | 605 | | | | |
| | EWPM | 3697034.183 | 605 | | | | |
| Corrected Total | EMLUW | 667.100 | 604 | | | | |
| | ENDW | 422325.795 | 604 | | | | |
| | EWPM | 414890.425 | 604 | | | | |

a Computed using alpha = .05

b R Squared = .055 (Adjusted R Squared = .051)

c R Squared = .079 (Adjusted R Squared = .076)

d R Squared = .050 (Adjusted R Squared = .047)

Table 10. Post Hoc Tests for EMLU, ENDW and EWPM as a Function of Age (Cross-sectional Dataset)

| Dependent Variable | (I) Age | (J) Age | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|--------------------|--------------|--------------|-----------------------|------------|--------|-------------------------|-------------|
| | | | | | | Upper Bound | Lower Bound |
| EMLUW | 5.45 to 6.45 | 6.48 to 7.48 | -.3309(*) | .11853 | .015 | -.6095 | -.0522 |
| | | 7.52 to 8.52 | -.6741(*) | .11564 | .000 | -.9459 | -.4023 |
| | 6.48 to 7.48 | 5.45 to 6.45 | .3309(*) | .11853 | .015 | .0522 | .6095 |
| | | 7.52 to 8.52 | -.3432(*) | .10376 | .003 | -.5871 | -.0993 |
| | 7.52 to 8.52 | 5.45 to 6.45 | .6741(*) | .11564 | .000 | .4023 | .9459 |
| ENDW | 5.45 to 6.45 | 6.48 to 7.48 | -.11.36(*) | 2.894 | .000 | -18.16 | -4.56 |
| | | 7.52 to 8.52 | -20.66(*) | 2.823 | .000 | -27.29 | -14.02 |
| | 6.48 to 7.48 | 5.45 to 6.45 | 11.36(*) | 2.894 | .000 | 4.56 | 18.16 |
| | | 7.52 to 8.52 | -9.30(*) | 2.534 | .001 | -15.25 | -3.34 |
| | 7.52 to 8.52 | 5.45 to 6.45 | 20.66(*) | 2.823 | .000 | 14.02 | 27.29 |
| EWPM | 5.45 to 6.45 | 6.48 to 7.48 | 9.30(*) | 2.534 | .001 | 3.34 | 15.25 |
| | | 7.52 to 8.52 | - | 2.89242 | .000 | -18.3598 | -4.7615 |
| | 6.48 to 7.48 | 5.45 to 6.45 | .4814 | 2.96467 | .986 | -6.4876 | 7.4504 |
| | | 7.52 to 8.52 | - | 2.59543 | .000 | -17.1803 | -4.9782 |
| | 7.52 to 8.52 | 5.45 to 6.45 | 11.5607(*) | 2.89242 | .000 | 4.7615 | 18.3598 |
| | 6.48 to 7.48 | 11.0792(*) | 2.59543 | .000 | 4.9782 | 17.1803 | |

Based on observed means.

* The mean difference is significant at the .05 level.

Table 11. Post Hoc Tests for EMLU, ENDW and EWPM as a Function of Grade
(Cross-sectional Dataset)

| Dependent Variable | (I) EGrade | (J) EGrade | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|--------------------|------------|------------|-----------------------|-------------|------|-------------------------|-------------|
| | | | Lower Bound | Upper Bound | | Lower Bound | Upper Bound |
| EMLUW | K | 1 | -.25 | .112 | .063 | -.51 | .01 |
| | | 2 | -.61(*) | .108 | .000 | -.87 | -.36 |
| | 1 | K | .25 | .112 | .063 | -.01 | .51 |
| | | 2 | -.36(*) | .095 | .000 | -.58 | -.14 |
| | 2 | K | .61(*) | .108 | .000 | .36 | .87 |
| | | 1 | .36(*) | .095 | .000 | .14 | .58 |
| ENDW | K | 1 | -8.85(*) | 2.773 | .004 | -15.37 | -2.34 |
| | | 2 | -18.89(*) | 2.692 | .000 | -25.22 | -12.57 |
| | 1 | K | 8.85(*) | 2.773 | .004 | 2.34 | 15.37 |
| | | 2 | -10.04(*) | 2.360 | .000 | -15.58 | -4.49 |
| | 2 | K | 18.89(*) | 2.692 | .000 | 12.57 | 25.22 |
| | | 1 | 10.04(*) | 2.360 | .000 | 4.49 | 15.58 |
| EWPM | K | 1 | 1.2686 | 2.79138 | .892 | -5.2898 | 7.8270 |
| | | 2 | - | 2.71008 | .000 | -17.4027 | -4.6679 |
| | 1 | K | 11.0353(*) | 2.79138 | .892 | -7.8270 | 5.2898 |
| | | 2 | - | 2.37590 | .000 | -17.8861 | -6.7217 |
| | 2 | K | 12.3039(*) | 2.71008 | .000 | 4.6679 | 17.4027 |
| | | 1 | 11.0353(*) | 2.37590 | .000 | 6.7217 | 17.8861 |

Based on observed means.

* The mean difference is significant at the .05 level.

Examination of Tables 12 and 13 suggests that mean Spanish MLU, NDW, and WPM (SMLU, SNDW, and SWPM) tended to increase as a function of age and grade. In order to examine whether the increases were significantly different across age and grade, two MANOVAs and a series of univariate ANOVAs were conducted. The results of the MANOVAs, using Wilk's Lambda indicated that there was significance for Age ($F(6, 1000) = 17.27, <.0001, \eta_p^2 = .09$) and Grade ($F(6, 1200) = 20.10, p < .001, \eta_p^2 = .09$). Partial eta squared values suggested that effect size was medium for age and grade. In order to ascertain if the effects of age and grade were consistent across the language variables a series of univariate ANOVAs was conducted. Results of the ANOVAs (Tables 14 and 15) indicated that age and grade were significant for all of the variables.

Age was significant for SMLU, $p = .000, \eta_p^2 = .11$, SNDW, $p = .000, \eta_p^2 = .11$, and SWPM, $p = .000, \eta_p^2 = .09$. Partial eta squared values suggested medium effect sizes for SMLU, SNDW, and SWPM. In order to further examine these differences, Tukey post hoc analyses were performed. The post hoc analyses indicated significant differences among some age groups, but not all. As can be seen in Table 16, for SMLU and SNDW there was significance between all of the groups, $p < .05$. For SWPM, however, the older age group had significantly higher means than the other two age groups (5.45 to 6.45). There was no significant difference between age groups 5.45 to 6.45 and 6.48 to 7.48.

Grade was significant for SMLU, $p = .000, \eta_p^2 = .11$, SNDW, $p = .000, \eta_p^2 = .12$, and SWPM, $p = .000, \eta_p^2 = .09$. Partial eta squared values suggested medium effect sizes for SMLU, SNDW, and SWPM. In order to further examine these

differences, Tukey post hoc analyses were performed. The post hoc analyses indicated significant differences between some grades, but not all. As can be seen in Table 17, the post hoc analyses demonstrated that for SMLU and SNDW, the means were significantly different between all of the grades, with increasing scores as grade became higher. For SWPM, the mean score for the second grade score was significantly higher than those of kindergarten and first grade. No other comparisons were significant.

In summary, the results of the analyses of the cross-sectional data indicate that MLU, NDW, and WPM for English and Spanish generally increased significantly as age and grade increased. It should be noted that compared to the other variables, all post hoc comparisons for NDW were significant. MLU was significant in all but one post hoc comparison, English grade, in which there was no significant difference between means for kindergarten and first grade. For WPM, there was significance only between the highest age group and grade level and their lower counterparts; there was no significance between the two lower groups for age and grade. Partial eta squared values for MLU and NDW were slightly higher than for WPM for age and grade in English and Spanish. Effect size was medium for all conditions except for EWPM for age and grade (small). These results indicate that based on cross-sectional data both age and grade are comparable indices of time, with small to medium effect size (Cohen, 1988). Partial eta squared values were slightly higher for age than grade for English and slightly higher for grade than age in Spanish. However, according to guidelines suggested by Cohen (1998), effect sizes were not substantially different between age and grade.

Table 12. Mean Scores of SMLU, SNDW, and SWPM as a Function of Age (Cross-sectional Dataset)

| Variable | Age group | M | SD | N |
|----------|--------------|-------|-------|-----|
| SMLUW | 5.45 to 6.45 | 5.30 | .83 | 123 |
| | 6.48 to 7.48 | 5.69 | .78 | 179 |
| | 7.52 to 8.52 | 6.02 | .78 | 203 |
| SNDW | 5.45 to 6.45 | 70.41 | 19.16 | 123 |
| | 6.48 to 7.48 | 79.33 | 21.66 | 179 |
| | 7.52 to 8.52 | 88.97 | 20.17 | 203 |
| S WPM | 5.45 to 6.45 | 65.45 | 20.37 | 123 |
| | 6.48 to 7.48 | 70.58 | 21.30 | 179 |
| | 7.52 to 8.52 | 81.76 | 22.99 | 203 |

Table 13. Mean Scores of SMLU, SNDW, and SWPM as a Function of Grade (Cross-sectional Dataset)

| Variable | Grade | M | SD | N |
|----------|-------|-------|-------|-----|
| SMLUW | K | 5.30 | .83 | 138 |
| | 1 | 5.61 | .80 | 215 |
| | 2 | 6.00 | .78 | 252 |
| SNDW | K | 69.98 | 19.02 | 138 |
| | 1 | 78.15 | 21.24 | 215 |
| | 2 | 88.44 | 19.62 | 252 |
| SWPM | K | 64.77 | 20.30 | 138 |
| | 1 | 69.52 | 21.25 | 215 |
| | 2 | 80.90 | 22.95 | 252 |

Table 14. Univariate ANOVA Results for SMLU, SNDW, and SWPM as a Function of Age (Cross-sectional Dataset)

| Source | Dependent Variable | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|-----------------|--------------------|-------------------------|-----|-------------|-----------|------|---------------------|
| Corrected Model | SMLUW | 39.708(a) | 2 | 19.854 | 31.798 | .000 | .112 |
| | SNDW | 27161.328(b) | 2 | 13580.664 | 32.398 | .000 | .114 |
| | SWPM | 23283.050(c) | 2 | 11641.525 | 24.541 | .000 | .089 |
| Intercept | SMLUW | 15522.917 | 1 | 15522.917 | 24861.576 | .000 | .980 |
| | SNDW | 3056536.847 | 1 | 3056536.847 | 7291.642 | .000 | .936 |
| | SWPM | 2544172.195 | 1 | 2544172.195 | 5363.244 | .000 | .914 |
| Age | SMLUW | 39.708 | 2 | 19.854 | 31.798 | .000 | .112 |
| | SNDW | 27161.328 | 2 | 13580.664 | 32.398 | .000 | .114 |
| | SWPM | 23283.050 | 2 | 11641.525 | 24.541 | .000 | .089 |
| Error | SMLUW | 313.436 | 502 | .624 | | | |
| | SNDW | 210430.165 | 502 | 419.184 | | | |
| | SWPM | 238134.679 | 502 | 474.372 | | | |
| Total | SMLUW | 16920.581 | 505 | | | | |
| | SNDW | 3553489.000 | 505 | | | | |
| | SWPM | 3013555.766 | 505 | | | | |
| Corrected Total | SMLUW | 353.143 | 504 | | | | |
| | SNDW | 237591.493 | 504 | | | | |
| | SWPM | 261417.729 | 504 | | | | |

a R Squared = .112 (Adjusted R Squared = .109)

b R Squared = .114 (Adjusted R Squared = .111)

c R Squared = .089 (Adjusted R Squared = .085)

Table 15. Univariate ANOVA Results for SMLU, SNDW, and SWPM as a Function of Grade (Cross-sectional Dataset)

| Source | Dependent Variable | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|-----------------|--------------------|-------------------------|-----|-------------|-----------|------|---------------------|
| Corrected Model | SMLUW | 45.938(b) | 2 | 22.969 | 36.044 | .000 | .107 |
| | SNDW | 32351.843(c) | 2 | 16175.922 | 40.117 | .000 | .118 |
| | SWPM | 27653.686(d) | 2 | 13826.843 | 29.170 | .000 | .088 |
| Intercept | SMLUW | 18020.783 | 1 | 18020.783 | 28279.230 | .000 | .979 |
| | SNDW | 3527350.646 | 1 | 3527350.646 | 8747.889 | .000 | .936 |
| | SWPM | 2918778.614 | 1 | 2918778.614 | 6157.659 | .000 | .911 |
| Grade | SMLUW | 45.938 | 2 | 22.969 | 36.044 | .000 | .107 |
| | SNDW | 32351.843 | 2 | 16175.922 | 40.117 | .000 | .118 |
| | SWPM | 27653.686 | 2 | 13826.843 | 29.170 | .000 | .088 |
| Error | SMLUW | 383.621 | 602 | .637 | | | |
| | SNDW | 242740.279 | 602 | 403.223 | | | |
| | SWPM | 285352.705 | 602 | 474.008 | | | |
| Total | SMLUW | 20094.218 | 605 | | | | |
| | SNDW | 4202650.000 | 605 | | | | |
| | SWPM | 3552794.748 | 605 | | | | |
| Corrected Total | SMLUW | 429.559 | 604 | | | | |
| | SNDW | 275092.122 | 604 | | | | |
| | SWPM | 313006.391 | 604 | | | | |

a Computed using alpha = .05

b R Squared = .107 (Adjusted R Squared = .104)

c R Squared = .118 (Adjusted R Squared = .115)

d R Squared = .088 (Adjusted R Squared = .085)

Table 16. Post Hoc Tests for SMLU, SNDW and SWPM as a Function of Age (Cross-sectional Dataset)

| Dependent Variable | (I) Age | (J) Age | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | | |
|--------------------|--------------|--------------|-----------------------|------------|---------|-------------------------|-------------|----------|
| | | | | | | Upper Bound | Lower Bound | |
| SMLUW | 5.45 to 6.45 | 6.48 to 7.48 | -.3829(*) | .09254 | .000 | -.6005 | -.1654 | |
| | | 7.52 to 8.52 | -.7159(*) | .09029 | .000 | -.9281 | -.5036 | |
| | 6.48 to 7.48 | 5.45 to 6.45 | .3829(*) | .09254 | .000 | .1654 | .6005 | |
| | | 7.52 to 8.52 | -.3329(*) | .08102 | .000 | -.5234 | -.1425 | |
| | 7.52 to 8.52 | 5.45 to 6.45 | .7159(*) | .09029 | .000 | .5036 | .9281 | |
| | | 6.48 to 7.48 | .3329(*) | .08102 | .000 | .1425 | .5234 | |
| | SNDW | 5.45 to 6.45 | 6.48 to 7.48 | -8.91(*) | 2.398 | .001 | -14.55 | -3.28 |
| | | | 7.52 to 8.52 | -18.55(*) | 2.339 | .000 | -24.05 | -13.05 |
| 6.48 to 7.48 | | 5.45 to 6.45 | 8.91(*) | 2.398 | .001 | 3.28 | 14.55 | |
| | | 7.52 to 8.52 | -9.64(*) | 2.099 | .000 | -14.57 | -4.70 | |
| 7.52 to 8.52 | | 5.45 to 6.45 | 18.55(*) | 2.339 | .000 | 13.05 | 24.05 | |
| | | 6.48 to 7.48 | 9.64(*) | 2.099 | .000 | 4.70 | 14.57 | |
| SWPM | | 5.45 to 6.45 | 6.48 to 7.48 | -5.1246 | 2.55084 | .111 | -11.1208 | .8717 |
| | | | 7.52 to 8.52 | - | 2.48867 | .000 | -22.1547 | -10.4546 |
| | 6.48 to 7.48 | 5.45 to 6.45 | 16.3047(*) | 5.1246 | 2.55084 | .111 | -.8717 | 11.1208 |
| | | 7.52 to 8.52 | - | 2.23314 | .000 | -16.4295 | -5.9307 | |
| | 7.52 to 8.52 | 5.45 to 6.45 | 11.1801(*) | 11.1801(*) | 2.48867 | .000 | 10.4546 | 22.1547 |
| | | 6.48 to 7.48 | 16.3047(*) | 2.48867 | .000 | 10.4546 | 22.1547 | |
| | | | 11.1801(*) | 2.23314 | .000 | 5.9307 | 16.4295 | |

Based on observed means.

* The mean difference is significant at the .05 level.

Table 17. Post Hoc Tests for SMLU, SNDW and SWPM as a Function of Grade
(Cross-sectional Dataset)

| Dependent Variable | (I) SGrade | (J) SGrade | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|--------------------|------------|------------|-----------------------|------------|------|-------------------------|-------------|
| | | | | | | Upper Bound | Lower Bound |
| SMLUW | K | 1 | -.3124(*) | .08707 | .001 | -.5169 | -.1078 |
| | | 2 | -.6969(*) | .08454 | .000 | -.8955 | -.4982 |
| | 1 | K | .3124(*) | .08707 | .001 | .1078 | .5169 |
| | | 2 | -.3845(*) | .07411 | .000 | -.5586 | -.2104 |
| | 2 | K | .6969(*) | .08454 | .000 | .4982 | .8955 |
| | | 1 | .3845(*) | .07411 | .000 | .2104 | .5586 |
| SNDW | K | 1 | -8.17(*) | 2.190 | .001 | -13.32 | -3.02 |
| | | 2 | -18.46(*) | 2.126 | .000 | -23.46 | -13.47 |
| | 1 | K | 8.17(*) | 2.190 | .001 | 3.02 | 13.32 |
| | | 2 | -10.29(*) | 1.864 | .000 | -14.67 | -5.91 |
| | 2 | K | 18.46(*) | 2.126 | .000 | 13.47 | 23.46 |
| | | 1 | 10.29(*) | 1.864 | .000 | 5.91 | 14.67 |
| SWPM | K | 1 | -4.7492 | 2.37477 | .113 | -10.3288 | .8303 |
| | | 2 | - | 2.30561 | .000 | -21.5453 | -10.7112 |
| | 1 | K | 16.1282(*) | 2.37477 | .113 | -.8303 | 10.3288 |
| | | 2 | - | 2.02130 | .000 | -16.1281 | -6.6299 |
| | 2 | K | 11.3790(*) | 2.30561 | .000 | 10.7112 | 21.5453 |
| | | 1 | 16.1282(*) | 2.02130 | .000 | 6.6299 | 16.1281 |

Based on observed means.

* The mean difference is significant at the .05 level.

Longitudinal Dataset

In order to examine whether there were significant differences across age and grade for MLU, NDW, and WPM, several repeated measures ANOVAs were conducted on data from the Longitudinal Dataset III. These were followed by Least Significant Differences (LSD) comparisons, for each variable, in order to determine which ages and grades were significantly different. A test of sphericity was first conducted to examine the equality of variance between the pairs of treatment conditions for repeated measurements (Davis, 2002). The results of these analyses indicated that the sphericity assumption was satisfied for all of the analyses in English ($p > .05$) and for most of the analyses in Spanish ($p > .05$). For the one occasion that there was a violation of the sphericity assumption, SMLU by Grade (Table 40), the corrected analyses using Greenhouse-Geisser corrections, were identical to those in which sphericity was assumed.

Effect size was interpreted using Cohen's guidelines (1988, Appendix A) for eta squared. It should be noted that eta squared (the equivalent of R squared) and partial eta squared values are slightly different (Becker, 1999, p. 3 - 4; Pallant, 2007), particularly with analyses that are more complex than a one-way ANOVA (Young, 1993). Studies involving within subject designs (i.e. repeated measures) might have larger effect size indices than those with between-subject designs (Young, 1993). Each language was examined separately. The results of the analyses of the English longitudinal data are presented first, followed by the analyses of the Spanish longitudinal data.

Examination of Tables 18 and 19 suggest that EMLU, ENDW, and EWPM tended to increase as a function of age and grade. The standard deviation differences were also worthy of mention. Generally, there was a trend for higher standard deviation values for the lower grades, which tended to decrease with higher grades. For ENDW, there was higher standard deviation for first grade ENDW as compared to kindergarten and second grades.

In order to examine whether the increases were significantly different across age and grade, separate Repeated Measures ANOVAs were conducted for each of the variables. The first set of Repeated Measures ANOVAs for English was conducted in order to determine whether EMLU, ENDW and EWPM were significantly different across age groups. The results of the first set of Repeated Measures ANOVAs (Tables 20 to 22) indicated that Age was significant for EMLU, $F(2, 70) = 55.07, p = .000, \eta_p^2 = .61$, ENDW, $F(2, 70) = 128.58, p = .000, \eta_p^2 = .79$, and EWPM, $F(2, 70) = 86.02, p = .000, \eta_p^2 = .71$. Partial eta squared values suggested that effect size was very large for all of the language variables. In order to further examine the differences, Least Significant Difference (LSD) comparisons were conducted to determine whether there were significant differences between pairwise comparisons for the language variables. LSD comparisons revealed that there were significant differences between all of the age groups, with higher age groups demonstrating higher scores. As can be seen in Tables 23, 24, and 25, there was significance for EMLU, $p < .001$, ENDW, $p < .05$, and EWPM, $p < .001$.

Table 18. Longitudinal Dataset III: English Language Variables
by Age (N=72)

| Variable | Age group | Mean | SD |
|-----------------|------------------|-------------|-----------|
| E MLUW | 5.2 to 6.2 | 5.65 | 1.16 |
| | 6.2 to 7.2 | 6.63 | .96 |
| | 7.2 to 8.2 | 7.12 | .83 |
| ENDW | 5.2 to 6.2 | 53.44 | 20.66 |
| | 6.2 to 7.2 | 84.32 | 23.96 |
| | 7.2 to 8.2 | 90.25 | 18.43 |
| EWPM | 5.2 to 6.2 | 60.84 | 23.39 |
| | 6.2 to 7.2 | 76.55 | 23.10 |
| | 7.2 to 8.2 | 90.39 | 20.50 |

Table 19. Longitudinal Dataset III: English Language Variables
by Grade (N=74)

| Variable | Grade | Mean | SD |
|-----------------|--------------|-------------|-----------|
| EMLUW | K | 5.65 | 1.15 |
| | 1 | 6.62 | .95 |
| | 2 | 7.11 | .82 |
| ENDW | K | 53.11 | 20.55 |
| | 1 | 84.23 | 23.64 |
| | 2 | 90.19 | 18.45 |
| EWPM | K | 60.76 | 23.13 |
| | 1 | 76.70 | 22.92 |
| | 2 | 90.70 | 20.47 |

Table 20. Repeated Measures ANOVA Results for EMLU as a Function of Age (Longitudinal Dataset III)

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|--------------------|--------------------|-------------------------|-----|-------------|--------|------|---------------------|
| EMLU by Age | Sphericity Assumed | 80.704 | 2 | 40.352 | 68.115 | .000 | .490 |
| Error(EMLU by Age) | Sphericity Assumed | 84.122 | 142 | .592 | | | |

a. Computed using alpha = .05

Table 21. Repeated Measures ANOVA Results for ENDW as a Function of Age (Longitudinal Dataset III)

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|--------------------|--------------------|-------------------------|-----|-------------|---------|------|---------------------|
| ENDW by Age | Sphericity Assumed | 56234.065 | 2 | 28117.032 | 136.060 | .000 | .657 |
| Error(ENDW by Age) | Sphericity Assumed | 29344.602 | 142 | 206.652 | | | |

Table 22. Repeated Measures ANOVA Results for EWPM as a Function of Age (Longitudinal Dataset III)

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|--------------------|--------------------|-------------------------|-----|-------------|--------|------|---------------------|
| EWPM by Age | Sphericity Assumed | 31478.500 | 2 | 15739.250 | 83.308 | .000 | .540 |
| Error(EWPM by Age) | Sphericity Assumed | 26827.892 | 142 | 188.929 | | | |

Table 23. Least Significant Difference Comparisons for EMLU as a Function of Age (Longitudinal Dataset III)

| (I) MLUbyage | (J) MLUbyage | Mean Difference (I-J) | Std. Error | Sig.(a) | 95% Confidence Interval for Difference(a) | |
|--------------|--------------|-----------------------|------------|---------|---|-------------|
| | | | | | Upper Bound | Lower Bound |
| 5.2 to 6.2 | 6.2 to 7.2 | -.977(*) | .121 | .000 | -1.219 | -.736 |
| | 7.2 to 8.2 | -1.471(*) | .143 | .000 | -1.756 | -1.186 |
| 6.2 to 7.2 | 5.2 to 6.2 | .977(*) | .121 | .000 | .736 | 1.219 |
| | 7.2 to 8.2 | -.494(*) | .120 | .000 | -.732 | -.255 |
| 7.2 to 8.2 | 5.2 to 6.2 | 1.471(*) | .143 | .000 | 1.186 | 1.756 |
| | 6.2 to 7.2 | .494(*) | .120 | .000 | .255 | .732 |

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Table 24. Least Significant Difference Comparisons for ENDW as a Function of Age (Longitudinal Dataset III)

| (I) ENDWbyage | (J) ENDWbyage | Mean Difference (I-J) | Std. Error | Sig.(a) | 95% Confidence Interval for Difference(a) | |
|---------------|---------------|-----------------------|------------|---------|---|-------------|
| | | | | | Upper Bound | Lower Bound |
| 5.2 to 6.2 | 6.2 to 7.2 | -30.875(*) | 2.231 | .000 | -35.324 | -26.426 |
| | 7.2 to 8.2 | -36.806(*) | 2.572 | .000 | -41.935 | -31.676 |
| 6.2 to 7.2 | 5.2 to 6.2 | 30.875(*) | 2.231 | .000 | 26.426 | 35.324 |
| | 7.2 to 8.2 | -5.931(*) | 2.372 | .015 | -10.660 | -1.201 |
| 7.2 to 8.2 | 5.2 to 6.2 | 36.806(*) | 2.572 | .000 | 31.676 | 41.935 |
| | 6.2 to 7.2 | 5.931(*) | 2.372 | .015 | 1.201 | 10.660 |

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Table 25. Least Significant Difference Comparisons for EWPM as a Function of Age (Longitudinal Dataset III)

| (I) EWPMbyage | (J) EWPMbyage | Mean Difference (I-J) | Std. Error | Sig.(a) | 95% Confidence Interval for Difference(a) | |
|---------------|---------------|-----------------------|------------|---------|---|-------------|
| | | | | | Upper Bound | Lower Bound |
| 5.2 to 6.2 | 6.2 to 7.2 | -15.708(*) | 2.344 | .000 | -20.381 | -11.034 |
| | 7.2 to 8.2 | -29.551(*) | 2.237 | .000 | -34.011 | -25.090 |
| 6.2 to 7.2 | 5.2 to 6.2 | 15.708(*) | 2.344 | .000 | 11.034 | 20.381 |
| | 7.2 to 8.2 | -13.843(*) | 2.290 | .000 | -18.410 | -9.276 |
| 7.2 to 8.2 | 5.2 to 6.2 | 29.551(*) | 2.237 | .000 | 25.090 | 34.011 |
| | 6.2 to 7.2 | 13.843(*) | 2.290 | .000 | 9.276 | 18.410 |

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

A second set of Repeated Measures ANOVAs for English was conducted in order to determine whether EMLU, ENDW and EWPM were significantly different across grade. The results of the second set of Repeated Measures ANOVAs (Tables 26 to 28) indicated that Grade was significant for EMLU, $F(2, 72) = 57.42, p = .000, \eta_p^2 = .62$, ENDW, $F(2, 72) = 136.92, p = .000, \eta_p^2 = .79$, and EWPM, $F(2, 72) = 91.67, p = .000, \eta_p^2 = .72$. Partial eta squared values suggested that effect sizes were very large for all of the language variables. In order to further examine the differences, LSD comparisons were conducted. LSD comparisons indicated that there were significant differences between all of the grades, with higher grades demonstrating higher scores. As can be seen in Tables 29 to 31, there was significance for EMLU, $p < .001$, ENDW, $p < .05$, and EWPM, $p < .001$.

Table 26. Repeated Measures ANOVA Results for EMLU as a Function of Grade (Longitudinal Dataset III)

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---------------------------------------|--------------------|-------------------------|-----|-------------|--------|------|---------------------|
| EMLU by Grade Error(EMLU by Grade) | Sphericity Assumed | 82.682 | 2 | 41.341 | 71.047 | .000 | .493 |
| | Sphericity Assumed | 84.955 | 146 | .582 | | | |

Table 27. Repeated Measures ANOVA Results for ENDW as a Function of Grade (Longitudinal Dataset III)

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---------------------------------------|--------------------|-------------------------|-----|-------------|---------|------|---------------------|
| ENDW by Grade Error(ENDW by Grade) | Sphericity Assumed | 58683.901 | 2 | 29341.950 | 143.968 | .000 | .664 |
| | Sphericity Assumed | 29756.099 | 146 | 203.809 | | | |

Table 28. Repeated Measures ANOVA Results for EWPM as a Function of Grade (Longitudinal Dataset III)

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|---------------------------------------|--------------------|-------------------------|-----|-------------|--------|------|---------------------|
| EWPM by Grade Error(EWPM by Grade) | Sphericity Assumed | 33219.253 | 2 | 16609.627 | 89.586 | .000 | .551 |
| | Sphericity Assumed | 27068.948 | 146 | 185.404 | | | |

Table 29. Least Significant Difference Comparisons for EMLU as a Function of Grade (Longitudinal Dataset III)

| (I) EMLUgrade | (J) EMLUgrade | Mean Difference (I-J) | Std. Error | Sig.(a) | 95% Confidence Interval for Difference(a) | |
|---------------|---------------|-----------------------|------------|---------|---|-------------|
| | | | | | Upper Bound | Lower Bound |
| K | 1 | -.973(*) | .118 | .000 | -1.208 | -.738 |
| | 2 | -1.469(*) | .140 | .000 | -1.748 | -1.191 |
| 1 | K | .973(*) | .118 | .000 | .738 | 1.208 |
| | 2 | -.496(*) | .117 | .000 | -.730 | -.263 |
| 2 | K | 1.469(*) | .140 | .000 | 1.191 | 1.748 |
| | 1 | .496(*) | .117 | .000 | .263 | .730 |

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Table 30. Least Significant Difference Comparisons for ENDW as a Function of Grade (Longitudinal Dataset III)

| (I) ENDW by grade | (J) ENDW by grade | Mean Difference (I-J) | Std. Error | Sig.(a) | 95% Confidence Interval for Difference(a) | |
|-------------------|-------------------|-----------------------|------------|---------|---|-------------|
| | | | | | Upper Bound | Lower Bound |
| K | 1 | -31.122(*) | 2.184 | .000 | -35.475 | -26.768 |
| | 2 | -37.081(*) | 2.515 | .000 | -42.092 | -32.070 |
| 1 | K | 31.122(*) | 2.184 | .000 | 26.768 | 35.475 |
| | 2 | -5.959(*) | 2.331 | .013 | -10.604 | -1.315 |
| 2 | K | 37.081(*) | 2.515 | .000 | 32.070 | 42.092 |
| | 1 | 5.959(*) | 2.331 | .013 | 1.315 | 10.604 |

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Table 31. Least Significant Difference Comparisons for EWPM as a Function of Grade (Longitudinal Dataset I)

| (I) EWPM by grade | (J) EWPM by grade | Mean Difference (I-J) | Std. Error | Sig.(a) | 95% Confidence Interval for Difference(a) | |
|-------------------|-------------------|-----------------------|------------|---------|---|-------------|
| | | | | | Upper Bound | Lower Bound |
| K | 1 | -15.944(*) | 2.287 | .000 | -20.502 | -11.385 |
| | 2 | -29.943(*) | 2.196 | .000 | -34.320 | -25.566 |
| 1 | K | 15.944(*) | 2.287 | .000 | 11.385 | 20.502 |
| | 2 | -13.999(*) | 2.231 | .000 | -18.445 | -9.553 |
| 1 | K | 29.943(*) | 2.196 | .000 | 25.566 | 34.320 |
| | 1 | 13.999(*) | 2.231 | .000 | 9.553 | 18.445 |

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Examination of Tables 32 and 33 suggest that mean SMLU and SWPM tended to increase as a function of age and grade. With regard to SNDW, there was an increase between kindergarten and first grade followed by a decrease between first and second grades. There was higher standard deviation in first grade as compared to kindergarten and second grades for SNDW, the same pattern noted for ENDW. This was different from the usual pattern of higher standard deviation values for the lower grades, which tended to decrease as grades became higher (for the SMLU and SWPM).

In order to examine whether the differences were significant across age and grade, separate repeated measures ANOVAs were conducted for each of the variables. The first set of Repeated Measures ANOVAs for Spanish was conducted to determine whether SMLU, SNDW and SWPM were significantly different across age groups. The results of the first set of Repeated Measures ANOVAs (Tables 34 to 36) indicated that Age was significant for SMLU, $F(2, 155) = 97.20, p = .000, \eta_p^2 = .56$, SNDW, $F(2, 155) = 242.86, p = .000, \eta_p^2 = .76$, and SWPM, $F(2, 155) = 39.19, p =$

.000, $\eta_p^2=.34$. Partial eta squared values suggested that effect sizes for SMLU and SNDW were very large, while the values for SWPM were large. In order to further examine the differences, LSD comparisons were conducted. The LSD comparisons indicated significant differences between some of the age groups. As can be seen in Table 37, there was significance between all of the age groups for SMLU, $p<.001$, with higher age groups demonstrating higher scores. For SNDW (Table 38), scores for the 6.2 to 7.2 age group was higher than those for the 5.1 to 6.1 age group, $p<.001$ and the 7.2 to 8.1 age group, $p<.05$. The 7.2 to 8.1 age group was higher than the 5.1 to 6.1 age group, $p<.001$. For SWPM (Table 39), the 6.2 to 7.2 and 7.2 to 8.1 age groups were significantly higher than the 5.1 to 6.1 age group, $p<.001$. There was no significant difference between the 6.2 to 7.2 and 7.2 to 8.1 age groups for SWPM.

Table 32. Longitudinal Dataset III: Spanish Language Variables by Age (N=157)

| Variable | Age group | M | SD |
|----------|------------|-------|-------|
| SMLUW | 5.1 to 6.1 | 4.88 | .85 |
| | 6.2 to 7.2 | 5.76 | .77 |
| | 7.2 to 8.1 | 6.08 | .83 |
| SNDW | 5.1 to 6.1 | 55.73 | 18.33 |
| | 6.2 to 7.2 | 90.62 | 20.56 |
| | 7.2 to 8.1 | 86.96 | 16.61 |
| S WPM | 5.1 to 6.1 | 57.38 | 20.21 |
| | 6.2 to 7.2 | 70.50 | 20.15 |
| | 7.2 to 8.1 | 73.63 | 19.89 |

Table 33. Longitudinal Dataset III: Spanish Language
Variables by Grade (N=163)

| Variable | Grade | Mean | SD |
|-----------------|--------------|-------------|-----------|
| SMLUW | K | 4.90 | .84 |
| | 1 | 5.76 | .77 |
| | 2 | 6.07 | .83 |
| SNDW | K | 56.11 | 18.37 |
| | 1 | 90.72 | 20.30 |
| | 2 | 87.13 | 16.65 |
| SWPM | K | 57.42 | 20.47 |
| | 1 | 70.45 | 20.03 |
| | 2 | 73.57 | 19.79 |

Table 34. Repeated Measures ANOVA Results for Spanish MLU as a Function of Age (Longitudinal Dataset III)

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|--------------------|--------------------|-------------------------|-----|-------------|---------|------|---------------------|
| SMLU by Age | Sphericity Assumed | 120.458 | 2 | 60.229 | 115.666 | .000 | .426 |
| Error(SMLU by Age) | Sphericity Assumed | 162.464 | 312 | .521 | | | |

Table 35. Repeated Measures ANOVA Results for Spanish NDW as a Function of Age (Longitudinal Dataset III)

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|--------------------|--------------------|-------------------------|-----|-------------|---------|------|---------------------|
| SNDW by Age | Sphericity Assumed | 115479.036 | 2 | 57739.518 | 247.379 | .000 | .613 |
| Error(SNDW by Age) | Sphericity Assumed | 72822.297 | 312 | 233.405 | | | |

Table 36. Repeated Measures ANOVA Results for Spanish WPM as a Function of Age (Longitudinal Dataset III)

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|--------------------|--------------------|-------------------------|-----|-------------|--------|------|---------------------|
| SWPM by Age | Sphericity Assumed | 23326.636 | 2 | 11663.318 | 40.491 | .000 | .206 |
| Error(SWPM by Age) | Sphericity Assumed | 89871.823 | 312 | 288.051 | | | |

Table 37. Least Significant Difference Comparisons for Spanish MLU as a Function of Age (Longitudinal Dataset III)

| (I) SMLU by age | (J) SMLU by age | Mean Difference (I-J) | Std. Error | Sig.(a) | 95% Confidence Interval for Difference(a) | |
|-----------------|-----------------|-----------------------|------------|---------|---|-------------|
| | | | | | Upper Bound | Lower Bound |
| 5.1 to 6.1 | 6.2 to 7.2 | -.880(*) | .079 | .000 | -1.037 | -.723 |
| | 7.2 to 8.1 | -1.195(*) | .089 | .000 | -1.370 | -1.020 |
| 6.2 to 7.2 | 5.1 to 6.1 | .880(*) | .079 | .000 | .723 | 1.037 |
| | 7.2 to 8.1 | -.315(*) | .076 | .000 | -.464 | -.166 |
| 7.2 to 8.1 | 5.1 to 6.1 | 1.195(*) | .089 | .000 | 1.020 | 1.370 |
| | 6.2 to 7.2 | .315(*) | .076 | .000 | .166 | .464 |

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Table 38. Least Significant Difference Comparisons for Spanish NDW as a Function of Age (Longitudinal Dataset III)

| (I) SNDW by age | (J) SNDW by age | Mean Difference (I-J) | Std. Error | Sig.(a) | 95% Confidence Interval for Difference(a) | |
|-----------------|-----------------|-----------------------|------------|---------|---|-------------|
| | | | | | Upper Bound | Lower Bound |
| 5.1 to 6.1 | 6.2 to 7.2 | -34.898(*) | 1.759 | .000 | -38.372 | -31.424 |
| | 7.2 to 8.1 | -31.229(*) | 1.694 | .000 | -34.576 | -27.883 |
| 6.2 to 7.2 | 5.1 to 6.1 | 34.898(*) | 1.759 | .000 | 31.424 | 38.372 |
| | 7.2 to 8.1 | 3.669(*) | 1.719 | .034 | .272 | 7.065 |
| 7.2 to 8.1 | 5.1 to 6.1 | 31.229(*) | 1.694 | .000 | 27.883 | 34.576 |
| | 6.2 to 7.2 | -3.669(*) | 1.719 | .034 | -7.065 | -.272 |

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Table 39. Least Significant Difference Comparisons for Spanish WPM as a Function of Age (Longitudinal Dataset III)

| (I) SWPMage | (J) SWPMage | Mean Difference (I-J) | Std. Error | Sig.(a) | 95% Confidence Interval for Difference(a) | |
|-------------|-------------|-----------------------|------------|---------|---|-------------|
| | | | | | Upper Bound | Lower Bound |
| 5.1 to 6.1 | 6.2 to 7.2 | -13.121(*) | 1.879 | .000 | -16.833 | -9.410 |
| | 7.2 to 8.1 | -16.242(*) | 1.962 | .000 | -20.118 | -12.367 |
| 6.2 to 7.2 | 5.1 to 6.1 | 13.121(*) | 1.879 | .000 | 9.410 | 16.833 |
| | 7.2 to 8.1 | -3.121 | 1.905 | .103 | -6.884 | .641 |
| 7.2 to 8.1 | 5.1 to 6.1 | 16.242(*) | 1.962 | .000 | 12.367 | 20.118 |
| | 6.2 to 7.2 | 3.121 | 1.905 | .103 | -.641 | 6.884 |

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

A second set of Repeated Measures ANOVAs was conducted to determine whether the Spanish measures (SMLU, SNDW and SWPM) were significantly different across grades. The results of the second set of Repeated Measures ANOVAs (Tables 40 to 42) indicated that Grade was significant for SMLU, $F(2, 161) = 100.21$, $p = .000$, $\eta_p^2 = .56$, SNDW, $F(2, 161) = 252.54$, $p = .000$, $\eta_p^2 = .76$, and SWPM, $F(2, 161) = 39.29$, $p = .000$, $\eta_p^2 = .33$. Partial eta squared values suggested that effect sizes for SMLU and SNDW were very large, while the values for SWPM were large. In order to further examine the differences, LSD comparisons were conducted. The LSD comparisons indicated significant differences between some of the grades. As can be seen in Table 43, there was significance between all of the grades for SMLU, $p < .001$, with higher grades demonstrating higher scores. For SNDW (Table 44), first grade was higher than kindergarten, $p < .001$ and second grade, $p < .05$. Second grade was higher than kindergarten, $p < .001$. For SWPM (Table 45), first and second grades were significantly higher than kindergarten, $p < .001$. There was no significant difference between first and second grades for SWPM.

Table 40. Repeated Measures ANOVA Results for SMLU as a Function of Grade (Longitudinal Dataset I)

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|----------------------|--------------------|-------------------------|---------|-------------|---------|------|---------------------|
| SMLU by grade | Sphericity Assumed | 122.378 | 2 | 61.189 | 120.038 | .000 | .426 |
| | Greenhouse-Geisser | 122.378 | 1.925 | 63.559 | 120.038 | .000 | .426 |
| Error(SMLU by grade) | Sphericity Assumed | 165.158 | 324 | .510 | | | |
| | Greenhouse-Geisser | 165.158 | 311.917 | .529 | | | |

Table 41. Repeated Measures ANOVA Results for SNDW as a Function of Grade (Longitudinal Dataset III)

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|----------------------|--------------------|-------------------------|-----|-------------|---------|------|---------------------|
| SNDW by grade | Sphericity Assumed | 118049.575 | 2 | 59024.787 | 257.356 | .000 | .614 |
| Error(SNDW by grade) | Sphericity Assumed | 74309.759 | 324 | 229.351 | | | |

Table 42. Repeated Measures ANOVA Results for SWPM as a Function of Grade (Longitudinal Dataset III)

| Source | | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|----------------------|--------------------|-------------------------|-----|-------------|--------|------|---------------------|
| SWPM by grade | Sphericity Assumed | 23939.532 | 2 | 11969.766 | 41.707 | .000 | .205 |
| Error(SWPM by grade) | Sphericity Assumed | 92987.136 | 324 | 286.997 | | | |

Table 43. Least Significant Difference Comparisons for SMLU as a Function of Grade (Longitudinal Dataset III)

| (I) SMLUgrade | (J) SMLUgrade | Mean Difference (I-J) | Std. Error | Sig.(a) | 95% Confidence Interval for Difference(a) | |
|---------------|---------------|-----------------------|------------|---------|---|-------------|
| | | | | | Upper Bound | Lower Bound |
| K | 1 | -.872(*) | .078 | .000 | -1.025 | -.719 |
| | 2 | -1.182(*) | .086 | .000 | -1.352 | -1.011 |
| 1 | K | .872(*) | .078 | .000 | .719 | 1.025 |
| | 2 | -.310(*) | .073 | .000 | -.454 | -.165 |
| 2 | K | 1.182(*) | .086 | .000 | 1.011 | 1.352 |
| | 1 | .310(*) | .073 | .000 | .165 | .454 |

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Table 44. Least Significant Difference Comparisons for SNDW as a Function of Grade (Longitudinal Dataset III)

| (I) SNDWgrade | (J) SNDWgrade | Mean Difference (I-J) | Std. Error | Sig.(a) | 95% Confidence Interval for Difference(a) | |
|---------------|---------------|-----------------------|------------|---------|---|-------------|
| | | | | | Upper Bound | Lower Bound |
| K | 1 | -34.607(*) | 1.702 | .000 | -37.968 | -31.246 |
| | 2 | -31.018(*) | 1.660 | .000 | -34.296 | -27.741 |
| 1 | K | 34.607(*) | 1.702 | .000 | 31.246 | 37.968 |
| | 2 | 3.589(*) | 1.671 | .033 | .290 | 6.888 |
| 2 | K | 31.018(*) | 1.660 | .000 | 27.741 | 34.296 |
| | 1 | -3.589(*) | 1.671 | .033 | -6.888 | -.290 |

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

Table 45. Least Significant Difference Comparisons for SWPM as a Function of Grade (Longitudinal Dataset III)

| (I) SWPMgrade | (J) SWPMgrade | Mean Difference (I-J) | Std. Error | Sig.(a) | 95% Confidence Interval for Difference(a) | |
|---------------|---------------|-----------------------|------------|---------|---|-------------|
| | | | | | Upper Bound | Lower Bound |
| K | 1 | -13.038(*) | 1.840 | .000 | -16.672 | -9.405 |
| | 2 | -16.153(*) | 1.946 | .000 | -19.995 | -12.311 |
| 1 | K | 13.038(*) | 1.840 | .000 | 9.405 | 16.672 |
| | 2 | -3.114 | 1.842 | .093 | -6.752 | .523 |
| 2 | K | 16.153(*) | 1.946 | .000 | 12.311 | 19.995 |
| | 1 | 3.114 | 1.842 | .093 | -.523 | 6.752 |

Based on estimated marginal means

* The mean difference is significant at the .05 level.

a Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).

In summary, the results of the analyses of the longitudinal data indicate that MLU, NDW, and WPM for English generally increased significantly as age and grade increased. In Spanish this significant increase was present for MLU and WPM, while for NDW there was a significant initial increase, followed by a decrease after first grade. Partial eta squared values were obtained as a measure of effect size. Using Cohen's guidelines (Cohen, 1988), partial eta squared values indicated large effect sizes for all of the English variables by age and grade. For Spanish, however, partial eta squared values were higher for MLU and NDW as compared to WPM. According to Cohen's guidelines, MLU and NDW for age and grade had very large effect sizes, while WPM for age and grade had large effect sizes. These results suggest that both age and grade are comparable indices of time, with large to very large effect size (Cohen, 1988) when examining the development of MLU, NDW, and WPM based on longitudinal data. However, the partial eta squared values were only slightly higher for grade than age in Spanish, with large to very large effect sizes across the language

variables. Additionally, partial eta squared values for MLU and NDW for age and grade in Spanish were much larger (very large effect size) compared to the value for WPM (large effect size).

Comparison Across Longitudinal Datasets

The results of the analyses of the cross-sectional and longitudinal data tend to suggest that age and grade are significant and comparable indices of time. Because the previous analyses had indicated no significant differences between age and grade and there was support in the developmental literature for using grade to examine changes in academic variables over time (e.g., Alexander & Martin, 2004, Morrison et al, 1997), grade was selected as the variable for the analyses comparing the three longitudinal datasets.

A set of analyses was conducted to address the question of whether the inclusion or exclusion of children with missing data or grade repeaters affects changes in the language measures. Longitudinal Dataset I (all children, including those who had data missing at one or more points and who repeated a grade), Longitudinal Dataset II (all children except for those who repeated a grade) and Longitudinal Dataset III (children who had data at all points and who did not repeat a grade) were used for these analyses. In order to examine whether means for MLU, NDW, and WPM were significantly different across longitudinal datasets, MANOVAs and a series of univariate ANOVAs were conducted. MANOVAs were conducted to determine whether there were significant differences for the datasets at each grade level. These were followed by a series of univariate ANOVAs to

determine for which language variables there was significance based on dataset.

Tukey post hoc analyses were conducted to ascertain between which datasets there was significance for each of the language variables. Interpretation of effect size was done using Cohen's guidelines (1988).

Examination of Table 46 suggests a tendency for higher EMLU, ENDW, and EWPM means in the Longitudinal Dataset III. In order to examine whether there were any significant differences, MANOVAs were conducted at each grade level (kindergarten, first grade and second grade) for the language variables. Results of the MANOVAs indicated that there were no significant differences between datasets at the kindergarten level for any of the language variables. However MANOVAs using Wilk's Lambda were significant at first grade, $F(6, 1630) = 3.6, p = .002, \eta_p^2 = .013$ and second grade $F(6, 1884) = 3.1, p = .005, \eta_p^2 = .010$. Partial eta squared values suggested very small effect sizes. In order to determine for which language variables there was significance, a series of univariate ANOVAs was conducted. Results indicated that there was significance at first grade (Table 47), dataset was significant for MLU, $p < .05, \eta_p^2 = .01$, NDW, $p < .001, \eta_p^2 = .02$, and WPM, $p < .05, \eta_p^2 = .02$. For second grade (Table 48) dataset was significant for MLU, $p < .05, \eta_p^2 = .01$, NDW, $p < .01, \eta_p^2 = .01$, and WPM, $p < .05, \eta_p^2 = .02$. Partial eta squared values suggested very small effect size for all the language variables (Cohen, 1988). To further examine the differences, post hoc tests were conducted. Results of the post hoc tests (Tables 49 and 50) indicated that means for Longitudinal Dataset III were significantly higher than those of Longitudinal Dataset I and Longitudinal Dataset II. for all of the language variables at first grade and second grade, $p < .05$. There were no differences

between the Longitudinal Dataset I and the Longitudinal Dataset II for any of the variables.

Table 46. Dataset Comparisons for Kindergarten, First Grade, and Second Grade for EMLU, ENDW, EWPM and SMLU, SNDW, and SWPM

| Grade | | Kindergarten | | | First Grade | | | Second | | | | |
|----------------|----------------------|--------------|-------|-----|-------------|-------|-------|--------|--|-------|-------|-----|
| English | | | | | | | | | | | | |
| Variable | Longitudinal Dataset | M | SD | N | | M | SD | N | | M | SD | N |
| EMLUW | I | 5.52 | 1.16 | 136 | | 6.14 | 1.32 | 386 | | 6.78 | .85 | 452 |
| | II | 5.56 | 1.14 | 127 | | 6.20 | 1.28 | 360 | | 6.83 | .80 | 422 |
| | III | 5.65 | 1.15 | 74 | | 6.62 | .95 | 74 | | 7.11 | .82 | 74 |
| ENDW | I | 52.11 | 21.74 | 136 | | 69.41 | 26.71 | 386 | | 78.51 | 23.73 | 452 |
| | II | 52.73 | 21.78 | 127 | | 70.28 | 26.54 | 360 | | 79.78 | 23.25 | 422 |
| | III | 53.11 | 20.55 | 74 | | 84.23 | 23.64 | 74 | | 90.19 | 18.45 | 74 |
| EWPM | I | 61.92 | 24.47 | 136 | | 65.41 | 24.93 | 386 | | 82.33 | 24.09 | 452 |
| | II | 61.83 | 24.10 | 127 | | 65.45 | 24.85 | 360 | | 82.68 | 23.75 | 422 |
| | III | 60.76 | 23.13 | 74 | | 76.70 | 22.92 | 74 | | 90.70 | 20.47 | 74 |
| Spanish | | | | | | | | | | | | |
| SMLUW | I | 4.91 | .87 | 257 | | 5.66 | .87 | 436 | | 5.95 | .8616 | 461 |
| | II | 4.90 | .84 | 238 | | 5.70 | .86 | 408 | | 5.99 | .85 | 428 |
| | III | 4.89 | .84 | 163 | | 5.76 | .77 | 163 | | 6.07 | .86 | 163 |
| SNDW | I | 56.99 | 18.56 | 257 | | 85.58 | 21.10 | 436 | | 85.02 | 17.02 | 461 |
| | II | 57.66 | 18.40 | 238 | | 86.27 | 20.97 | 408 | | 85.92 | 16.73 | 428 |
| | III | 56.11 | 18.37 | 163 | | 90.72 | 20.30 | 163 | | 87.13 | 16.65 | 163 |
| SWPM | I | 57.06 | 20.26 | 257 | | 65.44 | 19.77 | 436 | | 71.65 | 19.87 | 461 |
| | II | 57.64 | 20.42 | 238 | | 65.74 | 19.80 | 408 | | 72.20 | 19.87 | 428 |
| | III | 57.42 | 20.47 | 163 | | 70.45 | 20.03 | 163 | | 73.57 | 19.79 | 163 |

Table 47. Longitudinal Dataset Comparisons for EMLU, ENDW, and EWPM: First Grade

| Source | Dependent Variable | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|-----------------|--------------------|-------------------------|-----|-------------|-----------|------|---------------------|
| Corrected Model | EMLUW | 14.189(a) | 2 | 7.094 | 4.377 | .013 | .011 |
| | ENDW | 14099.517(b) | 2 | 7049.758 | 10.137 | .000 | .024 |
| | EWPM | 8550.811(c) | 2 | 4275.405 | 6.996 | .001 | .017 |
| Intercept | EMLUW | 19039.249 | 1 | 19039.249 | 11747.254 | .000 | .935 |
| | ENDW | 2655509.470 | 1 | 2655509.470 | 3818.412 | .000 | .824 |
| | EWPM | 2281657.657 | 1 | 2281657.657 | 3733.307 | .000 | .820 |
| Dataset | EMLUW | 14.189 | 2 | 7.094 | 4.377 | .013 | .011 |
| | ENDW | 14099.517 | 2 | 7049.758 | 10.137 | .000 | .024 |
| | EWPM | 8550.811 | 2 | 4275.405 | 6.996 | .001 | .017 |
| Error | EMLUW | 1324.145 | 817 | 1.621 | | | |
| | ENDW | 568181.521 | 817 | 695.449 | | | |
| | EWPM | 499319.921 | 817 | 611.163 | | | |
| Total | EMLUW | 32965.616 | 820 | | | | |
| | ENDW | 4731109.000 | 820 | | | | |
| | P20304W1EWPM | 4128380.390 | 820 | | | | |
| Corrected Total | P20304W1EMLUW | 1338.334 | 819 | | | | |
| | P20304W1ENDW | 582281.038 | 819 | | | | |
| | P20304W1EWPM | 507870.731 | 819 | | | | |

a R Squared = .011 (Adjusted R Squared = .008)

b R Squared = .024 (Adjusted R Squared = .022)

c R Squared = .017 (Adjusted R Squared = .014)

Table 48. Longitudinal Dataset Comparisons for EMLU, ENDW, and EWPM:
Second Grade

| Source | Dependent Variable | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
|-----------------|--------------------|-------------------------|-----|-------------|-----------|------|---------------------|
| Corrected Model | EMLUW | 7.119(a) | 2 | 3.559 | 5.267 | .005 | .011 |
| | ENDW | 8708.001(b) | 2 | 4354.001 | 8.127 | .000 | .017 |
| | EWPM | 4611.871(c) | 2 | 2305.936 | 4.113 | .017 | .009 |
| Intercept | EMLUW | 23738.782 | 1 | 23738.782 | 35124.733 | .000 | .974 |
| | ENDW | 3411947.976 | 1 | 3411947.976 | 6368.481 | .000 | .871 |
| | EWPM | 3613583.645 | 1 | 3613583.645 | 6445.975 | .000 | .872 |
| Dataset | EMLUW | 7.119 | 2 | 3.559 | 5.267 | .005 | .011 |
| | ENDW | 8708.001 | 2 | 4354.001 | 8.127 | .000 | .017 |
| | EWPM | 4611.871 | 2 | 2305.936 | 4.113 | .017 | .009 |
| Error | EMLUW | 638.671 | 945 | .676 | | | |
| | ENDW | 506288.821 | 945 | 535.755 | | | |
| | EWPM | 529762.615 | 945 | 560.595 | | | |
| Total | EMLUW | 44855.425 | 948 | | | | |
| | ENDW | 6580117.000 | 948 | | | | |
| | EWPM | 7087426.063 | 948 | | | | |
| Corrected Total | EMLUW | 645.790 | 947 | | | | |
| | ENDW | 514996.822 | 947 | | | | |
| | EWPM | 534374.487 | 947 | | | | |

a R Squared = .011 (Adjusted R Squared = .009)
b R Squared = .017 (Adjusted R Squared = .015)
c R Squared = .009 (Adjusted R Squared = .007)

Table 49. Post Hoc Tests for First grade Longitudinal Datasets: Comparisons by EMLU, ENDW and EWPM

| Dependent Variable | (I) Dataset | (J) Dataset | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|--------------------|-------------|-------------|-----------------------|------------|-------|-------------------------|-------------|
| | | | | | | Upper Bound | Lower Bound |
| EMLUW | I | II | -.0593 | .09328 | .801 | -.2783 | .1597 |
| | | III | -.4770(*) | .16156 | .009 | -.8563 | -.0976 |
| | II | I | .0593 | .09328 | .801 | -.1597 | .2783 |
| | | III | -.4177(*) | .16249 | .028 | -.7992 | -.0362 |
| | III | I | .4770(*) | .16156 | .009 | .0976 | .8563 |
| | | II | .4177(*) | .16249 | .028 | .0362 | .7992 |
| ENDW | I | II | -.87 | 1.932 | .893 | -5.41 | 3.66 |
| | | III | -14.82(*) | 3.347 | .000 | -22.68 | -6.96 |
| | II | I | .87 | 1.932 | .893 | -3.66 | 5.41 |
| | | III | -13.95(*) | 3.366 | .000 | -21.85 | -6.04 |
| | III | I | 14.82(*) | 3.347 | .000 | 6.96 | 22.68 |
| | | II | 13.95(*) | 3.366 | .000 | 6.04 | 21.85 |
| EWPM | I | II | -.0463 | 1.81135 | 1.000 | -4.2993 | 4.2068 |
| | | III | - | 3.13724 | .001 | -18.6583 | -3.9259 |
| | II | I | 11.2921(*) | 1.81135 | 1.000 | -4.2068 | 4.2993 |
| | | III | - | 3.15541 | .001 | -18.6547 | -3.8370 |
| | III | I | 11.2458(*) | 3.13724 | .001 | 3.9259 | 18.6583 |
| | | II | 11.2921(*) | 3.15541 | .001 | 3.8370 | 18.6547 |

Based on observed means.

* The mean difference is significant at the .05 level.

Table 50. Post Hoc Tests for Second Grade Longitudinal Datasets: Comparisons by EMLU, ENDW and EWPM

| Dependent Variable | (I) Dataset | (J) Dataset | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
|--------------------|-------------|-------------|-----------------------|------------|------|-------------------------|-------------|
| | | | | | | Upper Bound | Lower Bound |
| EMLUW | I | II | -.0508 | .05565 | .632 | -.1815 | .0798 |
| | | III | -.3345(*) | .10309 | .003 | -.5765 | -.0925 |
| | II | I | .0508 | .05565 | .632 | -.0798 | .1815 |
| | | III | -.2837(*) | .10361 | .017 | -.5269 | -.0405 |
| | III | I | .3345(*) | .10309 | .003 | .0925 | .5765 |
| | | II | .2837(*) | .10361 | .017 | .0405 | .5269 |
| ENDW | I | II | -1.27 | 1.567 | .696 | -4.95 | 2.41 |
| | | III | -11.68(*) | 2.903 | .000 | -18.49 | -4.87 |
| | II | I | 1.27 | 1.567 | .696 | -2.41 | 4.95 |
| | | III | -10.41(*) | 2.917 | .001 | -17.26 | -3.56 |
| | III | I | 11.68(*) | 2.903 | .000 | 4.87 | 18.49 |
| | | II | 10.41(*) | 2.917 | .001 | 3.56 | 17.26 |
| EWPM | I | II | -.3528 | 1.60271 | .974 | -4.1150 | 3.4094 |
| | | III | -8.3680(*) | 2.96915 | .014 | -15.3378 | -1.3982 |
| | II | I | .3528 | 1.60271 | .974 | -3.4094 | 4.1150 |
| | | III | -8.0152(*) | 2.98396 | .020 | -15.0198 | -1.0106 |
| | III | I | 8.3680(*) | 2.96915 | .014 | 1.3982 | 15.3378 |
| | | II | 8.0152(*) | 2.98396 | .020 | 1.0106 | 15.0198 |

Based on observed means.

* The mean difference is significant at the .05 level.

Examination of Table 46 suggests similar means for the Spanish longitudinal datasets for SMLU, SNDW, and SWPM. In order to examine whether there were any significant differences MANOVAs were conducted at each grade level (kindergarten, first grade and second grade) for the language variables. Results of the MANOVAs using Wilk's Lambda, indicated that there were no differences between datasets at the kindergarten, first grade and second grade levels for any of the language variables.

In summary, the results of the analyses of the longitudinal dataset comparisons indicate that for English at the kindergarten level there were no significant differences between datasets. However, there was significance at the first and second grade levels, with higher means for the language variables for Longitudinal Dataset III as compared to the other two datasets. There was no significant difference between Longitudinal Dataset I and Longitudinal Dataset II. Partial eta squared values suggested very small effect size. For Spanish, means were not significantly differences between any of the datasets.

CHAPTER 4

DISCUSSION

The purpose of this study was to examine the effect of using different approaches to obtaining normative data on the language skills of typically developing Spanish-speaking ELLs. Specific questions that were asked for this study were whether (a) grade or age variables should be used as an index of time for evaluating MLU, NDW and WPM, (b) norms vary as a function of using cross-sectional or longitudinal data, and (c) including or excluding children with missing data or those who have repeated a grade affects the language norms.

An alternative to standardized language tests that has been supported by the research literature is narrative language sampling. Narrative language sampling has been successfully used with speakers of English (e.g., Rice et al., 1998; Tilstra & McMaster, 2007) and other languages. However, adequate normative information on the language skills of ELLs is lacking due to small sample sizes, a focus on only one the languages spoken by ELLs (e.g., Muñoz et al., 2003) and the use of cross-sectional data (e.g., Muñoz et al., 2003; Miller et al., 2006). The SALT database (Miller & Iglesias, 2007) is the only existing large database examining both languages of Spanish-speaking ELLS. This database allows users to compare a child's performance to their age and grade peers on a variety of measures including MLU, NDW, and WPM. These three measures have been found to be useful in examining the narratives of ELLs since they are language neutral (e.g., Miller et al., 2006). Moreover, these measures have been proven to be reliable over time (Heilmann et al., 2008).

The datasets used for the current study, part of larger study examining the relationship between language skills and literacy, include both cross-sectional and longitudinal data obtained from children enrolled in kindergarten, first, and second grade. This project, called the Bilingual Language and Literacy Project (BLLP), consisted of kindergarten to third grade children enrolled in transitional bilingual for the cross-sectional sample, and kindergarten to second grade children enrolled in a variety of bilingual programs for the longitudinal sample. Only children in the transitional programs in kindergarten, first, and second grades were used for the current study. None of the children had been identified as having a disability. These datasets present a rare opportunity to study factors that might affect norming of language skills in ELLs.

Certain modifications were implemented to select the participants that were finally used for the study. For the cross-sectional dataset, of the original BLLP database of 928 children in kindergarten to third grade, 299 third grade children were eliminated in order to make the sample grade range comparable to that of the longitudinal dataset. An additional seven children were deleted because their ages appeared to have been incorrectly entered in the database (two years older than the oldest child in the final grade sample for a particular grade level). Two different datasets were subsequently created for the cross-sectional study in order to look at age and grade variables in English and Spanish. The sample selection for each language was performed separately in order to conserve the maximum number of participants. The grade dataset for English and Spanish samples consisted of 605 children. A total of 100 children were deleted from the grade dataset because they

did not conform to the strict age within grade requirements of having an age at the mean or six months above or below the mean of the children in the grade sample. There were 505 children for the English and Spanish age dataset. The number of children from the total sample eliminated for being out of range was 15 at kindergarten (2 percent of the total 605 sample; 11 percent of the kindergartners), 36 at first grade (6 percent of the total sample; 17 percent of the first graders), and 49 at second grade (eight percent of the total sample; 19 percent of the second graders). The implications regarding the percentage of out of range children on future language norming studies will be discussed later on.

From the longitudinal dataset of 679 kindergarten to second grade children, three different longitudinal datasets were created. The three datasets varied with respect to the inclusion or exclusion of repeaters and children who participated at all collection points. The Longitudinal Dataset I (679 children) was comprised of all the children, including children who had missing data at one or more of three data collection points (kindergarten, first grade, and second grade). Longitudinal Dataset II (639 children), consisted of all the children in Longitudinal Dataset I, except those who repeated a grade. The Longitudinal Dataset III consisted of children who had data at all three data collection points and did not repeat a grade. It should be noted that Longitudinal Dataset III was used to address the second research question (differences between cross-sectional and longitudinal datasets). Identical to the procedure conducted for the cross-sectional dataset, two different longitudinal datasets were constructed for Longitudinal Dataset III to examine age and grade variables in English and Spanish. The sample selection for each language was

performed separately in order to conserve the maximum number of participants. For the grade dataset, the English samples consisted of 74 children and the Spanish samples consisted of 163 children. The age dataset consisted of children at each grade level that were six months above or below the mean age for each grade. There were 72 children for the English age dataset and 157 children for the Spanish age dataset. It is important to note that the number of participants for the Spanish samples of Longitudinal Dataset III was greater than that of the English samples. This is the case because Spanish is the first language of the children sampled and many were not capable of producing narratives in English due to their limited English proficiency. Children out of range for age were two children (3%) for the English sample, and six children (4%) for the Spanish sample. For the longitudinal sample, the numbers of children deleted were two at kindergarten, two at first grade, and two at second grade for the English sample, and six at kindergarten, six at first grade, and six at second grade for the Spanish sample.

Age or Grade As An Index of Time?

Both age and grade have been used as indices of time in language development studies, with age being the preferred index used in standardized assessments. Some examples of standardized tests using age as an index of time are the CELF-4 English (Semel, Wiig, & Secord, 2003), the CELF-4 Spanish (Wiig, Secord, & Semel, 2005), the EWOPVT-SBE (Brownell, 2001), and the ROWPVT-SBE (Brownell, 2001). However, the literature on age and grade as an index of language development suggests that using norms based on age is not ideal once

children enter the educational system (Alexander & Martin, 2004). Once a child is exposed to the educational environment, factors associated with age level appear to lose their influence, resulting in a disparity between age and grade scores. Alexander and Martin (2004) recommended that once children enter the educational system, their development should be evaluated by grade norms.

The first question addressed in the present study focused on whether age or grade variables should be used as an index of time for examining MLU, NDW, and WPM. The results of the first set of analyses on the cross-sectional dataset indicated that age and grade values were similar and that there was an identical level of high statistical significance ($p = .000$) for age and grade for all of the language variables (MLU, NDW, and WPM). These results indicate that, at least in cross-sectional datasets, both age and grade variables are useful for accounting for differences in these language variables.

Although the partial eta squared values were slightly higher for age than grade for English and slightly higher for grade than age in Spanish, they were not substantially different based on the effect size guidelines suggested by Cohen (1998). Effect sizes for all of the variables for the cross-sectional study were medium except for English WPM for grade, which was small. These findings suggest that, at least in cross-sectional datasets, age and grade variables are comparable.

The results of the analyses of the longitudinal data indicate that age and grade scores were comparable, with a high level of statistical significance ($p = .000$) for MLU, NDW, and WPM. These results point to the validity of age and grade variables with longitudinal data in this area of language. With regard to practical significance,

partial eta squared values indicated large effect sizes (Cohen, 1998) for all of the English language variables by age and grade. For Spanish, partial eta squared values were higher (very large effect sizes) for MLU and NDW as compared to WPM (large effect size). Similar to the findings with the cross-sectional dataset, age and grade variables are also comparable in the longitudinal dataset.

The results of this study differ from what those found in the developmental educational literature, which reported significantly higher grade than age effects (Alexander & Martin, 2004; Cahan & Cohen, 1989; Crone & Whitehurst, 1999). For example, Cahan and Cohen (1989) used a quasi-experimental procedure that compared the performance on general ability tests of children who differed in chronological age and schooling to estimate the independent effects of age and schooling. Schooling was found to be the main factor causing an increase in intelligence test scores as a function of age. The results also suggested that schooling had a greater effect on verbal than on nonverbal tests. Crone and Whitehurst (1999) examined the emergent literacy and early reading skills of children longitudinally. They found that children who began school earlier than their same age peers, performed better than their peers on those skills; the effect of a year of school on literacy skills was 1.7 times greater than the gains related to age and 4.3 times greater than age on reading skills. Alexander and Martin (2004) evaluated the age within grade effect on reading mastery of first and second grade children. While significance was found for both age and grade, the effect of grade was twice the effect of age.

The comparable results of age and grade found for the current study differ from those found in the literature (e.g., Alexander & Martin, 2004; Cahan & Cohen,

1989; Cahan & Noyman, 2001; Crone & Whitehurst, 1999), possibly due to differences in methodology. An important difference between those studies (e.g., Alexander & Martin, 2004; Cahan & Cohen, 1989; Crone & Whitehurst, 1999) and the current one is the outcome measures studied. Cahan and Cohen (1989) used verbal cognitive ability tests, Crone and Whitehurst (1999) evaluated emergent literacy and early reading skills, and Alexander and Martin (2004) compared reading scores. The measures used in previous studies are traditional standardized instruments designed to assess academic skills, while the current study measured oral language skills. It is possible that as compared to academic skills, oral language skills are not substantially different when grade variables or age within grade variables are used.

Another difference between the current study and others in the literature is the criteria for age inclusion in the study. For the current study, children whose ages seemed to have been incorrectly entered in the database were deleted. Additionally, children who were out of range for grade were also excluded to create the age dataset. In the Alexander and Martin (2004) study, for example, there was also strict age within grade requirements that were imposed by the school system. However, in a small number of cases, there were exceptions made to the rules, resulting in a small percentage (2.3%) of the children being outside of the typical age range within grade. Alexander and Martin (2004) included the out of age range children in their sample. It is not certain that this is the reason for differing results between the two studies.

Currently, many educational tests include separate age and grade norms (e.g., the Kaufmann Educational Achievement Tests- Revised (KTEA-II) (Kaufman & Kaufman, 2004), the Wechsler Individual Achievement Test – Second Edition

(WIAT-II) (Weschler, 2001), and The Gray Oral Reading Test, Fourth Edition (GORT-4) (Weiderhold & Bryant, 2001). In contrast, developmental oral language tests tend to use age norms. However, some test developers are beginning to include separate grade norms as well (e.g., The PPVT IV (Dunn & Dunn, 2007)). In light of the importance of grade norms in the educational literature (e.g., Alexander & Martin, 2004; Cahan & Cohen, 1989) and the findings of the current study that grade norms significantly determine differences between language variables (specifically MLU, NDW, and WPM), it might be advantageous for researchers and clinicians to begin considering using grade as the index of time. If researchers and clinicians were to use grade as the index of time, it would provide them with the same time unit of measure for oral language and academic skills. This in turn will facilitate comparison of language skills and academic skills. Although not addressed in the present study, it is also possible that more academically related language skills might be more grade than age dependent. For example, although there are no significant differences between age and grade for NDW, it is possible that there would be noticeable differences if the task was to examine vocabulary typically learned in specific academic courses.

While grade is the recommended index, age should not be totally dismissed once children enter the academic situation. Age comparisons might be useful for identifying outliers (e.g., children who are outside the age range for their grade). In such cases, it would be helpful to determine whether a child who is too young or too old for the normative information pertaining to his grade level performs in a comparable manner to peers the same age in another grade (e.g., a first grader could be compared to a group of kindergartners of the same age). If the score received is

similar to those of same age peers, then the child would be performing age appropriately. The current study, like some others in the literature (e.g., Alexander & Martin), used an age within grade measure. The advantage of using age within grade scores is that outliers are eliminated. Subsequently this provides the opportunity to use alternative methods for examining the performance of outliers.

Cross-sectional or Longitudinal?

Almost all of the existing norms on language development are based on cross-sectional data. The emerging trend in language development (typical and atypical development) is to study longitudinal data (e.g., Annaz, Karmiloff-Smith, & Thomas, 2008; Dickinson & Tabors, 2001; Jarrold & Brock, 2004; Rice, 2004; Thomas, Annaz, Ansari, Scerif, Jarrold, & Karmiloff-Smith, 2009). The research literature highlights the importance of using longitudinal data to evaluate developmental phenomena, such as language and cognitive delays in developmental disorders (e.g., Annaz et al., 2008; Jarrold & Brock, 2004; Rice, 2004; Thomas et al., 2009). Some researchers recommend that cross-sectional studies be used for initial investigations, but that they should be followed by longitudinal studies (e.g., Annaz et al., 2008). The analysis of longitudinal studies is made more feasible and attractive to researchers with current data analysis procedures such as growth curve modeling (e.g., Duncan et al., 2006) and hierarchical linear modeling (e.g., Raudenbush & Bryk, 2001). Given this trend in the research literature, it is important to know whether any differences exist between using cross-sectional or longitudinal data when creating language norms.

For the current study, results of separate analyses of the cross-sectional dataset and Longitudinal Dataset III were used to compare results of the language variables MLU, NDW, and WPM for age and grade. Both datasets revealed significant statistical differences for all of the variables. However, there were some differences between results of the two datasets. First, effect size (partial eta squared values) for the longitudinal sample were much larger (mostly large; range large to very large) than those of the cross-sectional sample (mostly medium; range small to medium), indicating better practical significance for the longitudinal sample. Additionally, more instances of significant difference for age and grade were found with the longitudinal sample (in particular, English) as compared to the cross-sectional sample. For example, for all the longitudinal language variables in English, post hoc analyses showed significant differences for MLU, NDW, and WPM across all age groups and grades as compared to the cross-sectional sample for which there were differences between only some ages and grades for MLU and WPM. There was significance between all age groups and grades for NDW, however. In contrast to the English longitudinal results, the Spanish longitudinal results indicated differences between all age groups and grades only for NDW.

A noteworthy difference between the cross-sectional and longitudinal results is the direction of significance for NDW. In the cross-sectional results, there was a significant increase in NDW between all age groups and grade for English and Spanish. However, for the longitudinal results, this trend continued only for English. For Spanish NDW, there was a decrease between first and second grades after the initial increase between kindergarten and first grade. One explanation for the

discrepancy across datasets could be that NDW is more sensitive to changes in the children's use of their first and second language and this phenomenon was obscured with cross-sectional data. One possible explanation for the pattern difference in NDW for English and Spanish might be due to the child's acquisition of English lexical items. As ELLs progress through the transitional bilingual program, the amount of English that they learn and use increases. Although there should also be growth in their Spanish lexicon, this growth is likely to be reduced as the children become more English proficient. As children become more proficient in English, they are also more likely to begin including English words in their Spanish narratives. These English words used during the narration of "Spanish" narratives are not included in the Spanish NDW count, resulting in lower Spanish NDW score. For example, in the following example a child narrating the story in Spanish might use the word "eso/that" to refer to the owl. The subsequent year, the child might have learned the English word for the animal ("owl"), and rather than using the deictic "eso," while narrating the Spanish narrative would use "owl."

1. El **eso** está en el árbol (NDW Spanish = 6)
2. El **owl** está en el árbol (NDW Spanish = 5)

The net effect of including more English words in the "Spanish" narrative appears as if there is a reduction (attrition) or deceleration of NDW growth in Spanish. To what extent this intrusion of English is affecting the NDW and to what extent attrition or deceleration is occurring should be the focus of a future study.

The results of this study agree with the research literature regarding the advantage of using longitudinal data, particularly due to the presence of large effect

size values found for longitudinal results. Longitudinal data have been largely recognized as superior to cross-sectional data when studying developmental phenomena (e.g., Collier, 1992; Hedeker & Gibbons, 2006). Advantages of using longitudinal data (as compared to cross-sectional data) include greater statistical power (need for fewer subjects; variability is lessened since each subject acts as his/her own control), and the capability of studying change within individuals (Hedeker & Gibbons, 2006). The results of this study corroborate the literature on longitudinal research since effect size, an indication of practical significance for the longitudinal results is much higher than for the cross-sectional results. Since longitudinal data are more appropriate for chronicling development as compared to cross-sectional data, it suggests that the longitudinal results for NDW in this study might be more precise.

Inclusion or Exclusion of Participants With Missing

Data or Grade Repeaters?

Currently there are a number of advanced statistical methods that can be used with longitudinal data (e.g., Collier, 1992; Hedeker & Gibbons, 2006) to handle missing data and unique characteristics of the subjects in the norming sample. However, as a prerequisite to using these methods, it is important for researchers to have an understanding of how subject inclusion and exclusion might influence normative data. One issue that needs to be addressed is whether norms are different when including participants with or without missing data. Another issue is whether

norms including individuals who have repeated a grade differ substantially from those who do not include such individuals.

In order to answer the question of whether including or excluding children with missing data or those who have repeated a grade affects the language norms, the three longitudinal datasets were compared. Results demonstrated that the means for all the language variables for the Longitudinal Dataset III dataset (cases without missing data or containing no repeaters) were significantly higher than those for Longitudinal Dataset I (all children, including those with missing data and grade repeaters) and Longitudinal Dataset II (all the children except those who repeated a grade) for English at first and second grade levels. For all instances of significance, however, partial eta squared values were small, indicating little practical significance. For Spanish, there were no significant differences between datasets. The results of these analyses indicate that including or excluding subjects with missing data or repeaters does not significantly affect MLU, NDW, and WPM norms.

This lack of difference in datasets with and without subjects with missing data points, might be due to the use of a large dataset and the results might not be generalizable to small datasets. Often, when listwise deletion is used, causing a substantial decrease in the number of participants in longitudinal datasets, the accuracy of findings is compromised (e.g., Little & Rubin, 2002; Graham et al., 1994; Little & Schenker, 1995). This might be due to loss of power with data that are missing completely at random (MCAR) (participants who have missing data are not significantly different from those who have all data included, in terms of the variables under study) or missing at random (MAR) (with participants having missing data, the

missingness is unrelated to the value of a particular variable under study after controlling for another). However, if the data are not MCAR or MAR, a model must be created that takes into consideration the mechanism of missingness in those data. In this particular situation, it is recommended that a full longitudinal dataset be utilized because missingness can cause biases in parameter estimates. It is useful to use listwise deletion (deleting cases with missing information across data points) when there is a large amount of missing data in cases of MCAR or MAR, and the remaining cases have sufficient power to perform analyses. However, where the amount of power would be compromised when the missingness is MCAR or MCAR, or when missingness is based on variables under study, a method of preserving all the existing data must be used. Such methods include multiple imputation (Fichman & Cummings, 2003; Schafer & Graham, 2002; Wayman, 2003) or maximum likelihood estimation (Duncan et al., 2006; Schafer & Graham, 2002).

The United States Department of Education reports state that retention rates for Hispanics are lower than those of Blacks, but higher than those of Whites (Llagas & Snyder, 2003). The national percentage of repeaters is estimated to be seven to nine percent yearly (NCES, 1995). One reason for retention is low academic achievement (e.g. Brophy, 2006; Llagas & Snyder, 2003). However, for ELLs, retention due to low academic performance might be related to the children's limited English language proficiency, particularly in the early grades.

The results of the current study indicate that being retained does not negatively affect language measures. There were no differences in oral language measures (MLU, NDW, or WPM) for English or Spanish whether or not grade

repeaters were included in the longitudinal datasets. It is important to note that only 40 children (six percent of the children in Longitudinal Dataset I) were grade repeaters. This lack of difference with the inclusion of repeaters might be partially due to the low percentage of repeaters in the study. While for this study grade repeaters did not significantly affect results, future research should consider grade retention as a factor (particularly if the sample contains a large percentage of grade repeaters or subjects enrolled in higher grades). Studies have shown that the temporary improvement in achievement that occurs during the first year that the children repeat a grade (Hong & Raudenbush, 2005; Karweit, 1999) is short-lived and does not extend into subsequent grades. In effect, these children lag substantially behind their peers of similar age that had been promoted on measures of academic achievement. Moreover, studies have reported that grade repetition in kindergarten or other early grades result in comparably negative outcomes as grade repetition in higher grades (Hong & Raudenbush, 2005; Jimerson, 2001; Shepard & Smith, 1989).

The Language Measures, MLU, NDW, and WPM

The results of this study also provide further information on the value of using MLU, NDW, and WPM as measures of language development in ELLs. The results of this investigation substantiate the findings in the literature on language development that MLU (e.g., Bedore & Leonard, 1998; de Villiers & de Villiers, 1973; Miller & Chapman, 1981; Paul & Alforde, 1993), NDW (e.g., Thordardottir & Namazi, 2007; Watkins et al., 1995) and WPM (Tilstra and McMaster, 2007) are valid measures of language development.

Findings of the present study indicate that for all of the language variables of the English and Spanish cross-sectional analyses and the English longitudinal analyses, there was a statistically significant increase in means for all of the language variables as age and grade increased. For the Spanish longitudinal variables, however, there was a statistically significant increase for MLU and WPM as age and grade increased. However, for NDW, there was an increase between kindergarten and first grade followed by a decrease between first and second grades.

Contrary to results of the current study, Muñoz et al. (2003) found that NDW was not a sensitive measure of language development. There were substantial methodological differences between the two studies. For example, in the Muñoz et al. (2003) study, children were instructed to tell a story with pictorial support after first viewing pictures, while in the current study, a retell procedure was utilized (the children were told the story by the examiner and subsequently required to retell the story while looking at the pictures). Another difference is that the participants in the Muñoz et al. (2003) study were preschool children of exclusively low social economic backgrounds, whereas in the current study, the children were kindergarten to second graders, with no restrictions on socioeconomic background. Additionally, the Muñoz et al. (2003) involved a small sample size (24 children), while the current study consisted of large groups of children.

It is possible that the decrease in NDW scores following a period of increase demonstrated in the present study might be due to attrition (Anderson, 1999) or normal variation in the use of the first language of ELLs in the process of acquiring their second language. There was also a high standard deviation for NDW in first

grade as compared to the other grades, (indicating greater variability in the scores of participants). This increase in the amount of variability probably reflects a transition period in which the children are acquiring a substantial number of English lexemes and some, but not all, of these children are incorporating these English lexemes into their Spanish narratives. The decrease in variability in the second grade level indicates more similarity across children with respect to the number of different words they use in Spanish and English..

MLU, a measure of language complexity, has been a consistent and reliable measurement of children's language development (e.g., Bedore & Leonard, 1998; de Villiers & de Villiers, 1973; Miller & Chapman, 1981; Paul & Alforde, 1993). For this project, there was an increase in MLU with age and grade for both the cross-sectional and longitudinal analyses. For MLU, there was significance for all but one post hoc condition for age and grade. This shows that MLU is able to differentiate children based on age and grade for Spanish-speaking ELLs.

WPM, a measure of language fluency, has also been shown in the literature to be a good indicator of language development. In the literature, WPM has been found to be able to differentiate between children of different grades (Tilstra and McMaster, 2007) as well as between children with typical language development and those with language impairment (Scott & Windsor, 2000). The use of WPM was also supported in this study as a significant indicator of language development. However, it is important to note that it is not as robust as MLU and NDW. In this study, WPM increased with age and grade, although effect size was smaller than those of MLU and NDW. The results of analyses with these variables have important implications

for evaluation and intervention with ELLs. For example, the results suggest that it is important to determine which language variables are most ideal for evaluating language development and determining goals for intervention. Selection of variables that do not adequately distinguish between language skill levels based on age or grade could be problematic in making decisions regarding evaluation and intervention (for example, not differentiating between language difference and disorder).

In summary, the results of the present study indicate that certain variables must be taken into consideration when developing norms for narratives used with Spanish-speaking ELLs. Both age and grade variables are significant indices of time for evaluating MLU, NDW, and WPM. Regarding cross-sectional and longitudinal data, longitudinal data are clearly more appropriate for studying the development of language variables. However, cross-sectional data can provide useful preliminary information for later longitudinal investigations. When dealing with children who are not excessively outside the age range for their grade and who have been previously identified as typically developing, then including children who have repeated a grade might not significantly affect study results. Likewise, cases with missing data might not bear substantially on the development of certain language variables. However, this latter finding might not be generalizable to other longitudinal studies, specifically because each dataset might contain differing mechanisms of missingness, which must be adequately investigated in the process of conducting longitudinal analyses. Finally, the results confirm the findings from the majority of the research literature that MLU, NDW, and WPM are valid variables for studying narrative development.

Study Limitations

There are several limitations to the present study. First, more current methods designed for longitudinal analyses could have been used as part of the study analyses (e.g., growth curve modeling (Duncan et al., 2006)). One approach used in the literature is to impute missing values. This might have provided greater power for the analyses, possibly increasing the level of significance for the variables, particularly WPM. The second limitation concerns the generalizability of the study results to other populations of Spanish-speaking ELLs. The database used for this study was comprised exclusively of children from Mexican backgrounds. It would be beneficial if future databases included adequate representation of children from different Hispanic backgrounds (e.g., Puerto Rican, Cuban, or Dominican). Thirdly, the study included a large number of children from kindergarten to second grade. However, it would be informative to investigate how the significant effects of the language variables (MLU, NDW, and WPM) would be influenced with children's continued exposure to English. It is likely that the English skills of Spanish-speaking ELLs would surpass their Spanish skills due to continued exposure to the academic environment and less emphasis on native language skills (Spanish).

Future Research

Addressing important concerns about the variables to consider when developing normative information for language variables (MLU, NDW, and WPM) associated with narratives, is a critical step in normative studies of Spanish-speaking ELLs. In the current study, issues related to normative study have been restricted to

examining age and grade variables as indices of time, whether cross-sectional or longitudinal data should be used in providing normative data on language skills, and whether the inclusion or exclusion of children with missing data or grade repeats affects the language measures. Future studies should include the use of current methods for dealing with longitudinal data (e.g., multiple imputation and maximum likelihood estimation); examine the inclusion and exclusion of out of range children (e.g., how varying the percentage of out of range children can affect results) and different percentages of grade repeaters (e.g., to determine to what extent grade repetition affects oral language skills); study age and grade effects in higher age and grade groups; and investigate the differential and cross-linguistic effects of English and Spanish on the developmental language measures.

Regarding the use of more current statistical methods for dealing with longitudinal data, it might be advantageous to look at data that include cases with missing items. This would give a more complete account of language development in that it will use all the available data efficiently. For many studies, it is more appropriate to use all the data, since deleting participants with missing data can seriously affect the results of longitudinal analyses (e.g., Little & Rubin, 2002; Graham et al., 1994; Twisk, 2003). Regarding out of range children, it would be useful to determine how varying the percentage of out of range children can affect results with oral language variables. Likewise, it would be informative to determine the effect of the inclusion of different percentages of grade repeaters. This information would help to more effectively establish the variables to control for when studying oral language skills.

This study showed that age and grade variables are comparable when examining MLU, NDW, and WPM in kindergarten, first and second grade children. However, it would be helpful to ascertain how the effects of age and grade are manifested in these language skills of older children. With respect to MLU, NDW, and WPM, future studies should make comparisons with the available developmental information on these variables for other populations (e.g., monolingual children in English and Spanish and children from various Hispanic backgrounds). Additionally, it would help to study the differences as well as interactions between the English and Spanish with respect to the different language variables (particularly NDW) because it would shed more light on the developmental course of these language variables. Concerning NDW, it would be informative to compare total NDW (conceptually scored) and individual language NDW (ENDW and SNDW).

REFERENCES CITED

- Allen, D. V., & Bliss, L. S. (1987). Concurrent validity of two language screening tests. *Journal of Communication Disorders, 20*, 305-317.
- Allen, M.S., Kertoy, M.K., Sherblom, J.C., & Pettit, J. M. (1994). Children's narrative productions: A comparison of personal event and fictional stories. *Applied Psycholinguistics, 15*, 149-176.
- Alexander, J. R.M. & Martin, F. (2004). The end of the reading age: grade and age effects in early schooling. *Journal of School Psychology, 42*(5), 403-416.
- Anderson, R. (1999). Impact of first language loss on grammar in a bilingual child. *Communication Disorders Quarterly, 21*, 4-16.
- Annaz, D., Karmiloff-Smith, A., & Thomas, M. S., C. (2008). The importance of tracing developmental trajectories for clinical child neuropsychology. In J. Reed & J. Warner-Rogers (Eds.), *Child neuropsychology: Concepts, theory and practice* (pp. 7-18). Oxford, United Kingdom: Wiley-Blackwell
- Aram, D., Morris, R., & Hall, N. (1993). Clinical research congruence in identifying children with specific language impairment. *Journal of Speech and Hearing Research, 36*, 580-591
- Battle, D. (2002). Language development and disorders in culturally and linguistically diverse children. In D. Bernstein & E. Tiegerman-Farber (Eds.), *Language and communication disorders in children* (pp. 354-386). Boston: Allyn Bacon.
- Becker, L.A. (1998). Measures of effect size: Strength of association. Retrieved June 9, 2009 from http://web.uccs.edu/lbecker/SPSS/glm_effectsize.htm . .
- Bedore, L.M. & Leonard (1998). Specific language impairment and grammatical morphology: a discriminant function analysis. *Journal of Speech, Language, and Hearing Research, 41*, 1185-1192.
- Bedore, L. M. & Peña, E.D. (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *The International Journal of Bilingual Education and Bilingualism, 11* (1), 1-16.

- Berman, R., & Slobin, D. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, NJ: Erlbaum.
- Bishop, D. V. M. & Edmundson, A (1987). Language impaired four year olds: Distinguishing transient from persistent impairment. *Journal of Speech and Hearing Disorders*, 52, 156-173.
- Blake, J., Quartaro, G., & Onorati, S. (1993). Evaluating quantitative measures of grammatical complexity in spontaneous speech samples. *Journal of Child Language*, 20, 139-152.
- Boudreau, D. M. & Hedberg, N. L. (1999). A comparison of early literacy skills in children with specific language impairment and their typically developing peers. *American Journal of Speech-Language Pathology*, 8, 249-260.
- Brice, A. E. (2002). *The Hispanic child*. Boston: Allyn & Bacon.
- Brice, A., & Montgomery, J. (1996). Adolescent pragmatic skills: A comparison of Latino students in ESL and speech-language programs. *Language, Speech, and Hearing Services in Schools*, 27, 68-81.
- Brophy, J. (2006). *Grade repetition*. The International Institute for Educational Planning (IIEP) and The International Academy of Education (IAE).
- Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.
- Brownell, R. (2001). *Expressive One-Word Picture Vocabulary Test (EOWPVT)*. San Antonio, TX: Pearson Education.
- Brownell, R. (2001). *Expressive One-Word Picture Vocabulary Test – Bilingual Edition (EOWPVT-SBE)*. San Antonio, TX: Pearson Education.
- Brownell, R. (2001). *Receptive One-Word Picture Vocabulary Test – Bilingual Edition (ROWPVT-SBE)*. San Antonio, TX: Pearson Education, Inc.
- Brownell, R. (2001). *Receptive One-Word Picture Vocabulary Test (ROWPVT)*. San Antonio, TX: Pearson Education.
- Cahan, S. & Cohen, N. (1989). Age versus schooling effects on intelligence. *Child Development*, 60, 1239-1249
- Cahan, S. & Noyman, A. (2001). The Kaufman Ability Battery for Children mental processing scale: A valid measure of 'pure' intelligence. *Educational and Psychology Measurement*, 61, 827-840.

- Cheng, L. (1991). *Assessing Asian language performance*. Oceanside, CA: Academic Communication Associates.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Collier, V. P. (1992). A synthesis of studies examining long-term language minority student data on academic achievement. *Bilingual Research Journal*, 16 (1 & 2), 187-212.
- Comeau, L., Genesee, F. & Lapaquette, L (2003). The modeling hypothesis and child bilingual codemixing. *The International Journal of Bilingualism*, 7(2), 113-126.
- Crais, E., & Lorch, N. (1994). Oral narratives in school-age children. *Topics in Language Disorders*, 14(3), 13-28.
- Crone, D.A. & Whitehurst, G.J. (1999). Age and schooling effects on emergent literacy and reading skills. *Journal of Educational Psychology*, 91, 604-614.
- Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*. New York, NY: Springer-Verlag.
- Deno, S. L. (2003). Developments in curriculum based measurement. *Journal of Special Education*, 37(3), 184-192.
- Deno, S. L., Mirkin, P. K., & Chaing, B. (1982). Identifying valid measures of reading. *Exceptional Children*, 49(1), 36-45.
- de Villiers, J. G. & de Villiers, P.A. (1973). A cross-sectional study of the acquisition of grammatical morphemes in child speech. *Journal of Psycholinguistic Research*, 2(3), 267-278.
- Dickinson, O.K. & Tabors, P. O. (Eds.) (2001). *Beginning literacy with language: Young children learning at home and school*. Baltimore: Paul H. Brookes.
- Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An Introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd ed.) . Mahwah, NJ: Lawrence Erlbaum Associates.
- Dunn, L. M. & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4)*. San Antonio, TX: Pearson Education, Inc.
- Dunn, L. M. Lugo, D. E., & Padilla, E. R. & Dunn, L. M. (1986). *Test de vocabulario en imágenes Peabody*. Pearson Education, Inc. Minneapolis, MN: Pearson Education, Inc.

- Dunn, M., Flax, J., Sliwinski, M., & Aram, D. (1996). The use of spontaneous language measures as criteria for identifying children with specific language impairment: An attempt to reconcile clinical and research incongruence. *Journal of Speech and Hearing Research, 39*, 643-654.
- Erickson, J. G. & Iglesias, A. (1986). Assessment of communication disorders in non-English proficiency children. In O. Taylor (Ed.). *Nature of communication disorders in culturally and linguistically diverse populations* (pp.181-217). San Diego, C.A: College Hill.
- Fazio, B., Naremore, R., & Connell, P. (1996). Tracking children from poverty at risk for specific language impairment: A 3-year longitudinal study. *Journal of Speech and Hearing Research, 39*, 611-624.
- Fichman, M. & Cummings, J. N. (2003). Multiple imputation for missing data: Making the most of what you know. *Organizational Research Methods, 6* (3), 282-308.
- Francis, D., Carlson, C., Fletcher, J., Foorman, B., Goldenberg, C., Vaughn, S., et al. (2005). Oracy/literacy development of Spanish-speaking children: A multi-level program of research on language minority children and the instruction, school and community contexts, and interventions that influence their academic outcomes. *Perspectives, 8-12*.
- Fuchs, L. S. (2004). The past, present, and future curriculum-based measurement research. *School Psychology Review, 33*(2), 188-192.
- Fuchs, L. S. Fuchs, D., & Speece, D. (2002). Treatment validity as a unifying construct for identifying learning disabilities. *Learning Disability Quarterly, 25*(1), 33-46.
- Gavin, W. J., Klee, T., & Membrino, I. (1993). Differentiating specific language impairment from normal language development using grammatical analysis. *Clinical Linguistics & Phonetics, 7* (3), 191-206.
- Genesee, F. & Nicoladis, E. (2007). Bilingual first language acquisition (p. 324-342). In E. Hoff & M. Shatz (Eds.). *Handbook of Language Development*. Harrisonburg, VA: Wiley-Blackwell.
- Goldstein, B.A. (2004). Bilingual language development and disorders: Introduction and overview. In B.A. Goldstein (Ed.). *Bilingual language development and disorders in Spanish-English speakers* (pp. 3-19). Baltimore, MD: Paul Brookes.

- Graham, J. W., Hofer, S. M. & Piccinin, A. M. (1994). Analysis of missing data in drug prevention research. In L. M. Collings & L. A. Seitz (Eds.). *Advances in data analysis for prevention intervention research*, 142. Rockville, MD: National Institute on Drug Abuse.
- Guasti, M. T. (2004). *Language acquisition: The growth of grammar*. Cambridge, MA: MIT Press.
- Gutierrez-Clellen, V. F., Restrepo, M. A., Bedore, L., Peña, E. & Anderson, R. (2000). Language sample analysis in Spanish-speaking children: Methodological considerations. . *Language, Speech and Hearing Services in Schools*, 31, 88-98.
- Heaton, R. K., & Marcotte, T. D. (2000). Clinical and neuropsychological tests and assessment techniques. In F. Boller, J. Grafman, & G. Rizzolatti (Eds.). *Handbook of neuropsychology* (2nd ed.). St. Louis, MO: Elsevier.
- Hedeker, D. & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: John Wiley & Sons.
- Heilmann, J., Miller, J., Iglesias, I., Fabiano-Smith, L., Nockerts, A., & Andriacchi, K. D. (2008). Narrative transcription accuracy and reliability in two languages. *Topics in Language Disorders*, 28(2).
- Heubert, J. P., & Hauser, R. M. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.
- Hewitt, L. E., Hammer, C.S., Yont, K. M., & Tomblin, J.B. (2005). Language sampling for kindergarten children with and without SLI: mean length of utterance, IPSYN, and NDW. *Journal of Communication Disorders*, 38, 197-213.
- Hieke, A. (1981). A content-processing view of hesitation phenomena. *Language and Speech*, 24, 147-160.
- Holowka, S., Brosseau-Lapr e, F., Petitto, L. A. (2002). Semantic and conceptual knowledge underlying bilingual babies' first signs and words. *Language Learning*, 52:2, 205-262.
- Hong, G., & Raudenbush, S. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, 27, 205-224.

- Hughes, D. McGillivray, L., & Schmidek, M. (1997). *Guide to narrative language: Procedures for assessment*. Eau Claire, WI: Thinking Publications.
- Hunt, K. W. (1965). *Grammatical structures written at three grade levels*. (Research Report No. 3). Champaign, IL: National Council of Teachers of English.
- Jarrold, C., & Brock, J. (2004). To match or not to match? Methodological issues in autism-related research. *Journal of Autism and Developmental Disorders*, *34*, 81-86.
- Jimerson, S. (2001). Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review*, *30*, 420-437.
- Justice, L. M., Bowles, R. P., Kaderavek, J. N., Ukrainetz, T. A., Eisenberg, S. L., & Gillam, R. B. (2006). The index of narrative microstructure: A clinical tool for analyzing school-age children's narrative performances. *American Journal of Speech-Language Pathology*, *15*, 177-191.
- Kaderavek, J. N. & Sulzby, E. (2000). Narrative production by children with and without specific language impairment: Oral narratives and emergent readings. *Journal of Speech, Language, and Hearing Research*, *43*, 34-49.
- Karweit, N. (1999). *Grade retention: Prevalence, timing, and effects* (Report No. 33). Baltimore: Center for Research on the Education of Students Placed at Risk, Johns Hopkins University.
- Kaufman, A. S. & Kaufman, N. L. (2004). *Kaufman Test of Educational Achievement, Second Edition (KTEA-II)*. Circle Pines, MN: American Guidance Service
- Kayser, H., & Restrepo, M. (1995). Language samples: Elicitation and analysis. In H. Kayser (Ed.), *Bilingual speech-language pathology: An Hispanic focus* (pp. 265-286). San Diego, CA: Singular.
- Klee, T. (1992). Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders*, *12*, 28-41.
- Klee, T., Schaffer, M., May, S., Membrino, I., & Mougey, K. (1989). A comparison of the age-MLU relation in normal and specifically language-impaired preschool children. *Journal of Speech and Hearing Disorders*, *54*, 226-233.
- Klee, T. Stokes, S., Wong, A., Fletcher, P. & Gavin, W. (2004). Utterance length and lexical diversity in Cantonese-speaking children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, *47*, 1396-1410.

- Kraemer, H. C., Yesavage, J.A., Taylor, J.L., & Kupfer, D. (2000). How can we learn about developmental processes from cross-sectional studies, or can we? *American Journal of Psychiatry*, 157(2), 163-171.
- Laing, S. P., & Kamhi, A. (2003). Alternative assessment of language and literacy in culturally and linguistically diverse populations. *Language, Speech, and Hearing Services in Schools*, 34, 44-55.
- Langdon, H. W. (1992). Speech and language assessment of LEP/bilingual Hispanic students. In H. W. Langdon & L. L. Cheng (Eds.). *Hispanic children and adults with communication disorders* (pp. 201-271). Gaithersburg, MD: Aspen
- Langdon, H. W. (2008). *Assessment and intervention for communication disorders in culturally and linguistically diverse populations*. Clifton Park, NY: Thomson Delmar Learning.
- Leadholm, B., & Miller, J. (1992). *Language sample analysis: The Wisconsin guide*. Madison, WI: Wisconsin Department of Public Instruction.
- Liles, B. Z. (1993). Narrative discourse in children with language disorders and children with normal language: A critical review of the literature. *Journal of Speech and Hearing Research*, 36, 868-882.
- Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data*. New York: John Wiley & Sons.
- Little, R. J. A. & Schenker, N. (1995). Missing data. In G. Arminger, CC. Clogg, & M. E. Sobel. (Eds). *Handbook of statistical modeling for the social and behavioral sciences*, pp 39-75. New York: Plenum.
- Llagas, C. & Snyder, T. D. (2003). Status and trends in the education of Hispanics. NCES 2003-008. U.S. Department of Education, National Center for Education Statistics.
- Loban, W. (1976), *Language development: Kindergarten through grade twelve* (Research Report 18). Urbana, IL: National Council of Teachers of English.
- MacLachlan, B., & Chapman R. (1988). Communication breakdowns in normal and language-learning disabled children's conversation and narration. *Journal of Speech and Hearing Disorders*, 53, 2-7.
- Martin, R., Foels, P., Clanton, G., & Moon, K. (2004). Season of birth is related to child retention rates, achievement, and rate of diagnosis of specific LD. *Journal of Learning Disabilities*, 37, 307-317.

- Mattes, L., & Omark, D. (1991). *Speech and language assessment for the bilingual handicapped*. Oceanside, CA: Academic Communication Associates.
- Merrell, A., & Plante, E. (1997). Norm-referenced test interpretation I the diagnostic process. *Language, Speech, and Hearing Services in Schools, 28*, 50-58.
- Miller, J. F. (1991). *Assessing language production in children: Experimental procedures*. Baltimore: University Park Press.
- Miller, J. F., & Chapman, R. S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech and Hearing Research, 24*, 154-161.
- Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., & Francis, J. (2006). Oral language and reading in bilingual children. *Learning Disabilities Research & Practice, 21(1)*, 30-43.
- Miller, J. & Iglesias, A. (2007). Systematic Analysis of English and Spanish Language Transcripts (Research Version 9).
- Morrison, F. J. Griffith, E.M. & Alberts, D. M. (1997). Nature-nurture in the classroom: Entrance age, school readiness, and learning in children. *Developmental Psychology, 33*, 254-262.
- Muñoz, M. L., Gillam, R. B., Peña, E.D., & Gulley -Faehle, A. (2003). Measures of language development in fictional narratives of Latino children. *Language, Speech, and Hearing Services in Schools, 34*, 332-342.
- National Clearinghouse for English Language Acquisition and Language Instruction in Educational Programs (NCELA) (2006). NCELA FAQ: How many school-age English language learners (ELLs) are there in the U.S? Available at <http://www.ncela.gwu.edu/expert/gaq/01leps.html>. Retrieved 10/15/08.
- National Clearinghouse for English Language Acquisition and Language Instruction in Educational Programs (NCELA) (2008). Survey of the states' limited English proficient students and available educational programs and services 200-2001 Summary Report (Kinder, 2002). Available at <http://www.ncela.gwu.edu/expert/fastfaq/4.html>. Retrieved 10/15/08.
- Owens, R. E. (1999). *Language disorders: A functional approach to assessment and intervention*. Boston, MA: Allyn & Bacon.
- Pallant, J. F. (2007). *SPSS Survival manual: A step by step guide to data analysis using SPSS*. Sydney : Allen & Unwin.

- Paul, R. (2001). *Language disorders from infancy to adolescence: Assessment and intervention (2ndnd ed)*. St. Louis, MO: Mosby
- Paul, R. & Alforde, S. (1993). Grammatical morpheme acquisition in 4-year-olds with normal, impaired, and late-developing language. *Journal of Speech and Hearing Research, 36*, 1271-275.
- Paul, R. & Smith, R. (1993). Narrative skills in 4-year-olds with normal, impaired, and late-developing language. *Journal of Speech and Hearing Research, 36*, 592-598
- Pearson, B. (2002). Narrative competence among monolingual and bilingual schoolchildren in Miami. In D. Oller & R. Eilers (Eds.), *Language and literacy in bilingual children* (pp. 135-174). Clevedon, UK: Multilingual Matters.
- Pearson, B. Z., & Fernández, S., C. (1994). Patterns of interaction in the lexical growth in two languages of bilingual infants and toddlers. *Language Learning, 44*, 617-653.
- Pearson, B.Z., Fernández, S. & Oller, D.K. (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language and Learning 43* (1), 93-120.
- Pearson, B.Z., Fernández, S.C., & Oller, D.K. (1995). Cross-language synonyms in the lexicons of bilingual infants: One language or two? *Journal of Child Language, 22* (2), 345-368.
- Peña, E. D., Bedore, L. M., & Zlatic-Giunta, R. (2002). Category generation performance of young bilingual children: The influence of condition, category, and language. *Journal of Speech, Language, and Hearing Research, 41*, 938-947.
- Peña, E. D, Spaulding, T. J. & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech-Language Pathology, 15*, 247-254.
- Pew Hispanic Center (2008). Latinos account for half of U.S. population growth since 2000. Retrieved on October 8, 2008, from <http://pewhispanic.org/reports/report.php?ReportID=96>
- Plante, E. & Vance, R. (1994). Selection of preschool language tests: A databased approach. *Language, Speech, and Hearing Services in Schools, 25*, 15-24.
- Plante, E., & Vance, R. (1995). Diagnostic accuracy of two tests of preschool language. *American Journal of Speech-Language Pathology, 4*, 70-76.

- Ramos, M. & Ramos, J. (2007). *The Test of Early Language Development-Third Edition: Spanish (TELD-3: S)*. Oceanside, CA: Academic Communication Associates, Inc.
- Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd. Ed)*. Thousand Oaks, CA: Sage Publications.
- Reese, L., Gallimore, R. & Guthrie, D. (2005). Reading trajectories of immigrant Latino students in transitional bilingual programs. *Bilingual Research Journal*, 29 (3), 679- 697.
- Rosanksy, E. J. (1976). Methods and morphemes in second language acquisition research. *Language Learning*, 26(2), 409-425.
- Rice, M. L. (2004). Growth models of developmental language disorders. In M. L. Rice & S. F. Warren (Eds.), *Developmental language disorders: From phenotypes to etiologies* (pp. 207-240). Mahwah, NJ: Erlbaum.
- Rice, M. L. Wexler, K., & Hershberger, S. (1998). Tense over time: The longitudinal course of tense acquisition in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 41, 1412-1431.
- Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes*, 14, 423-441.
- Rojas, R., Pereira, H., & Iglesias, A. (2000, November). *Utterance segmentation of Spanish language samples*. Paper presented at the American Speech-Language-Hearing Association Convention, Washington, DC.
- Saenz, T. I. & Huer, M. B. (2008). Testing strategies involving least biased assessment of bilingual children. *Communication Disorders Quarterly*, 24, 184-193.
- Schafer, J. L. & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7 (2), 147-177.
- Scott, C. M. & Windsor, J.(2000). General language performance measures in spoken and written narrative and expository discourse of school-age children with language learning disabilities. *Journal of Speech, Language, and Hearing Research*, 43, 324-339.
- Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals, Fourth Edition*. San Antonio, Tx: Pearson Education, Inc.
- Semel, Wiig, & Secord (1997). *Clinical Evaluation of Language Fundamentals – Third Edition Spanish (CELF-3 Spanish)*. San Antonio, TX: Pearson Education, Inc.

- Shepard, L., & Smith, M. (Eds.). (1989). *Flunking grades: Research and policies on retention*. London: Falmer.
- Switzer, J. & Gruber, C. P. (1992). *Norris Educational Achievement Test*. Los Angeles, CA: Western Psychological Services.
- Szaflarski, J. P. Schmithorst, V. J., Altaye, M., Byars, A.W. Ret, J., Plante, E., & Holland, S.K. (2006). A longitudinal study of language development in children age 5-11. *Ann. Neurol.*, 59(5), 796-807.
- Thomas, M.S.C, Annaz, D., Ansari, D., Scerif, G, Jarrold, C., & Karmiloff-Smith, A. (2009). Using developmental trajectories to understand developmental disorders. *Journal of Speech, Language, and Hearing Research*, 52, 336-358.
- Thordardottir, E. (1998). Mean length of utterance and other language sample measures in early Icelandic. *First Language*, 18(52), 001-32.
- Thordardottir, E. & Namazi, M. (2007). Specific language impairment in French-speaking children: Beyond grammatical morphology. *Journal of Speech, Language, and Hearing Research*, 50, 698-715.
- Tilstra, J. & McMaster, K. (2007). Productivity, fluency, and grammaticality measures from narratives: Potential indicators of language proficiency. *Communication Disorders Quarterly*, 29:1, 43-53.
- Twisk, J.W.R. (2003). *Applied longitudinal data analysis for epidemiology: A practical guide*. New York, NY: Cambridge.
- Uriate, M., Lavan, N. Agusti, N. & Karp, F. (2009). English learners in Boston public schools: Enrollment and educational outcomes of native Spanish speakers. Boston, MA: Gaston Institute for Latino Community Development and Public Policy.
- U.S. Census Bureau (1993). *Population Division and Housing and Household Economic Statistics Division*. Current Population Reports, Series P20-475, *The Hispanic Population in the United States: March 1993*. Retrieved August 4, 2009.
- U. S. Census Bureau (2000). The Census 2000 Summary File 3. factfinder.census.gov Retrieved August 4, 2009.
- U. S. Census Bureau (2007).
<http://www.census.gov/population/www/socdemo/school/cps2007.html>
 Retrieved August 1, 2009

- U. S. Census Bureau (2008). http://www.census.gov/Press-Release/www/releases/archives/facts_for_features_special_editions/012245.html Retrieved on October 8, 2008, from http://www.census.gov/Press-Release/www/releases/archives/facts_for_features_special_editions/012245.html
- U. S. Department of Education Report (2009). <http://www.ed.gov/fund/grant/find/edlite-forecast.html> Retrieved August 3, 2009.
- Van der Lely, H.K.J. (1997). Narrative discourse in grammatical specific language impaired children: A modular language deficit? *Journal of Child Language*, 24, 221-256.
- Washington, J. (1996). Issues in assessing the language abilities of African American children. In A. Kamhi, K. Pollock, & J. Harris (Eds.). *Communication development and disorders in African American children: Research, assessment, and intervention* (pp. 35-54). Baltimore: Brookes.
- Watkins, R. V., Kelly, D. J. & Harbers, H. M. & Hollis, W. (1995). Measuring children's lexical diversity: Differentiating typical and impaired language learners. *Journal of Speech and Hearing Research*, 38, 1349-1355.
- Wayman, J. C. (2003). Multiple imputation for missing data: What it is and how can I use it? Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL.
- Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C.A. (2007). Literature synthesis on curriculum-based measurement in reading. *Journal of Special Education*, 41(2), 85-120.
- Weiderhold, J. L. & Bryant, B. R. (2001). *Gray Oral Reading Test- Fourth Edition. (GORT-4)*. San Antonio, TX: Pearson.
- Weschler, D. (2001). *Weschler Individual Achievement Test- Second Edition. (WIAT-II)*. San Antonio, TX: Pearson.
- Wiig, Secord, & Semel (2005). *Clinical Evaluation of Language Fundamentals – Fourth Edition Spanish (CELF-4 Spanish)*. San Antonio, TX: Pearson Education, Inc.
- Woodcock, R. W. & Johnson, M. V. (1989). *The Woodcock Johnson Psycho-Educational Battery-revised (WJ-R)*. Chicago: Riverside
- Young, M.A. (1993). Supplementing tests of statistical significance: Variation accounted for. *Journal of Speech and Hearing Research*, 36, 644-656.

Zimmerman, I. L., Steiner, V.G., & Pond, R. E. (2002). *Preschool Language Scale-Fourth Edition (PLS-4) English Edition*. San Antonio, TX: Pearson Education, Inc Edition

Zimmerman, I. L., Steiner, V.G., & Pond, R. E. (2002). *Preschool Language Scale-Fourth Edition (PLS-4) Spanish Edition*. San Antonio, TX: Pearson Education, Inc.

APPENDIX A
MISSING DATA PATTERNS

Missing Data Patterns for English Narratives (3 data points): Longitudinal Dataset I

| Missingness Patterns | 1 | Wave 2 | 3 | Pattern Frequency |
|-------------------------|---|-----------|---|----------------------|
| 1 | X | X | X | 80 |
| 2 | X | X | | 12 |
| 3 | X | | X | 14 |
| 4 | X | | | 30 |
| 5 | | X | X | 215 |
| 6 | | X | | 79 |
| 7 | | | X | 143 |

Missing Data Patterns for Spanish Narratives (3 data points): Longitudinal Dataset I

| Missingness Patterns | 1 | Wave 2 | 3 | Pattern Frequency |
|-------------------------|---|-----------|---|----------------------|
| 1 | X | X | X | 177 |
| 2 | X | X | | 33 |
| 3 | X | | X | 16 |
| 4 | X | | | 31 |
| 5 | | X | X | 178 |
| 6 | | X | | 48 |
| 7 | | | X | 90 |

APPENDIX B

BILINGUAL SPANISH/ENGLISH STORY RETELL STORY SCRIPTS

English script for *Frog, Where Are You?* by Mercer Mayer, 1969.

Page Script

- 1** There once was a boy who had a dog and a pet frog. He kept the frog in a large jar in his bedroom.
- 2** One night while he and his dog were sleeping, the frog climbed out of the jar. He jumped out of an open window.
- 3** When the boy and the dog woke up the next morning, they saw that the jar was empty.
- 4** The boy looked everywhere for the frog. The dog looked for the frog too. When the dog tried to look in the jar, he got his head stuck.
- 5** The boy called out the open window, "Frog, where are you?" The dog leaned out the window with the jar still stuck on his head.
- 6** The jar was so heavy that the dog fell out of the window headfirst!
- 7** The boy picked up the dog to make sure he was ok. The dog wasn't hurt but the jar was smashed.
- 8 - 9** The boy and the dog looked outside for the frog. The boy called for the frog.
- 10** He called down a hole in the ground while the dog barked at some bees in a beehive.
- 11** A gopher popped out of the hole and bit the boy on right on his nose. Meanwhile, the dog was still bothering the bees, jumping up on the tree and barking at them.
- 12** The beehive fell down and all of the bees flew out. The bees were angry at the dog for ruining their home.
- 13** The boy wasn't paying any attention to the dog. He had noticed a large hole in a tree. So he climbed up the tree and called down the hole.
- 14** All of a sudden an owl swooped out of the hole and knocked the boy to the ground.
- 15** he dog ran past the boy as fast as he could because the bees were chasing him.
- 16** The owl chased the boy all the way to a large rock.
- 17** The boy climbed up on the rock and called again for his frog. He held onto some branches so he wouldn't fall.
- 18** But the branches weren't really branches! They were deer antlers. The deer picked up the boy on his head.
- 19** The deer started running with the boy still on his head. The dog ran along too. They were getting close to a cliff.
- 20-21** The deer stopped suddenly and the boy and the dog fell over the edge of the cliff.
- 22** There was a pond below the cliff. They landed with a splash right on top of one another.
- 23** They heard a familiar sound.
- 24** The boy told the dog to be very quiet.
- 25** They crept up and looked behind a big log.
- 26** There they found the boy's pet frog. He had a mother frog with him.
- 27** They had some baby frogs and one of them jumped towards the boy.
- 28-29** The baby frog liked the boy and wanted to be his new pet. The boy and the dog were happy to have a new pet frog to take home. As they walked away the boy waved and said "goodbye" to his old frog and his family.

Spanish script for *Frog, Where Are You?* by Mercer Mayer, 1969.

Página Papel

- 1** Había un niño quien tenía un perro y una rana. El tenía la rana en su cuarto en un jarro grande a su rana.
- 2** Una noche cuando el niño y su perro estaban durmiendo, la rana se escapó del jarro. La rana se salió por una ventana abierta.
- 3** Cuando el niño y el perro se despertaron la siguiente mañana, vieron que el jarro estaba vacío.
- 4** El niño buscó en todas partes a la rana. Aún adentro de sus botas. El perro también buscó a la rana. Cuando el perro trató de mirar adentro del jarro y no podía sacar la cabeza.
- 5** El niño empezó a llamar desde la ventana abierta: "Rana, ¿Dónde estás?". El perro se asomó a la ventana con el jarro todavía en la cabeza.
- 6** ¡El jarro estaba tan pesado que hizo que el perro se cayera de cabeza por la ventana!
- 7** El niño fue a ver como estaba el perro. El perro no estaba herido, pero el jarro se rompió.
- 8–9** El niño y el perro buscaron a la rana afuera de la casa. El niño llamó a la rana.
- 10** El niño llamaba a la rana en un hoyo que estaba en la tierra, mientras que el perro le ladraba a unas abejas en su panal.
- 11** Una ardilla salió de su hueco y mordió la nariz del niño por molestarla. Mientras tanto, el perro seguía molestando a las abejas, brincaba hacia el árbol y les ladraba.
- 12** El panal de abejas se cayó y las abejas salieron volando. Las abejas estaban enojadas con el perro.
- 13** El niño no prestó ninguna atención al perro. El vió un hueco grande en un árbol y quería ver si su rana se escondía allí. Así que trepó el árbol y llamó a la rana en el hueco para ver si estaba.
- 14** De repente un buho salió del hueco y lanzó al niño al suelo. El buho lo vió fijamente y le dijo que se fuera.
- 15** El perro pasó al niño corriendo tan rápido como pudo porque las abejas lo perseguían.
- 16** El buho persiguió al niño hasta una piedra grande.
- 17** El niño se encaramó en la piedra y llamó otra vez a la rana. Se agarró a unas ramas para no caerse de la piedra.
- 18** ¡Pero las ramas no eran ramas reales! Eran los cuernos de un venado. El venado le evantó al niño con su cabeza.
- 19** Y el venado empezó a correr con el niño que estaba todavía en su cabeza. El perro también corrió al lado del venado. Se acercaron a un precipicio.
- 20–21** El venado se paró de pronto y el niño y el perro se cayeron por el precipicio.
- 22** Había un estanque debajo del precipicio. Aterrizaron en el estanque uno encima del otro.
- 23** Oyeron un sonido que conocían.
- 24** El niño le dijo al perro que se callara.
- 25** Los dos se acercaron con cuidado y miraron detrás de un tronco de un árbol.
- 26** Allí encontraron a la rana del niño. Había con él una rana mamá también.
- 27** Ellos tenían algunas ranitas bebés y una de ellas saltó hacia el niño.
- 28–29** La ranita quería mucho al niño y quería ser su nueva mascota. El niño y el perro estaban felices de tener una nueva rana y llevarla a casa. Cuando se iban, el niño dijo adiós a la que fue su rana y también a su familia.

APPENDIX C
CALCULATION OF EFFECT SIZE

Calculation of effect size

<http://wev.uccs.edu/lbecker/Psy590/escal3.htm>.

Cohen's guidelines for eta squared (for group comparisons)

| Size | Eta squared (% of variance explained) | Cohen's d (standard deviation units) |
|--------|--|---|
| Small | .01 or 1% | .2 |
| Medium | .06 or 6% | .5 |
| Large | .138 or 13.8% | .8 |

A comparison of effect size values (from

http://web.uccs.edu/lbecker/SPSS/glm_effectsize.htm)

| Effect | η^2 | η_p^2 |
|----------------|----------------------------|------------------------------|
| Drive | .039 | .068 |
| Reward | .184 | .253 |
| Reward * Drive | .236 | .304 |