VALIDATING THE USE OF *D*:

MEASURING LEXICAL DIVERSITY IN LOW-INCOME CHILDREN

by

STEPHANIE ALICE LAI

(Under the Direction of Paula Schwanenflugel)

ABSTRACT

Children from low socioeconomic (SES) families are exposed fewer words compared to children from middle-income families, and as a consequence, often score lower on standardized vocabulary assessments (Hart & Risley, 1995) such as the Expressive Vocabulary Test (EVT). Words that are acquired as part of one's vocabulary are influenced by cultural experiences. An alternative measure, *D*, has been proposed to assess lexical diversity by comparing the number of unique words with the total number of words in a language or writing sample (Malvern, Richards, Chipere, & Durán, 2004). *D* provides an alternative measure to vocabulary deployment that is not linked to a child's knowledge of specific vocabulary words. The primary purpose of the current study is to validate *D* as a useful measure for lexical diversity in at-risk, low-income, predominantly African American children. *D* was validated using Kane's argument-based approach to validity (1992). Five assumptions were proposed to validate *D* as a measure of lexical diversity and are grounded in research regarding the validation of standardized vocabulary assessments in multicultural populations.

Based on the five assumptions, the findings from this study provide some evidence in support of *D* as a valid measure for evaluating lexical diversity in low-income children who are predominantly African American. *D* was found to be somewhat related to expressive vocabulary;

the relationship was weak and therefore suggests that *D* measures an aspect of vocabulary that is related to but different than expressive vocabulary. Further, there were no differences in *D* between African American and non-African American children. This suggests that *D* may overcome the racial-bias exhibited in standardized assessments. Previous research in standardized vocabulary tests found that these measures can be racially-biased with low-income children performing lower than the standardized mean and African American children performing even lower within that group. The evidence collected in this study provides support that *D* is generalizable across races and SES, and could be a useful supplement to standardized measures of vocabulary. The use of standardized and nonstandardized measures together could more effectively screen and detect early language deficits in order to determine language ability and guide intervention efforts.

INDEX WORDS: lexical diversity, d, validation, expressive vocabulary, kindergarten, low-income

VALIDATING THE USE OF *D*:

MEASURING LEXICAL DIVERSITY IN LOW-INCOME CHILDREN

by

STEPHANIE ALICE LAI

B.S., The University of Florida, 2006

M.A., The University of Florida, 2008

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2014

VALIDATING THE USE OF *D*:

MEASURING LEXICAL DIVERSITY IN LOW-INCOME CHILDREN

by

STEPHANIE ALICE LAI

Major Professor:	Paula Schwanenflugel
Committee:	Stacey Neuharth-Pritchett
	Nancy Knapp
	Liang Chen

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
August 2014

DEDICATION

Let parents bequeath to their children not riches, but the spirit of reverence. –Plato

I dedicate this piece of work to my Mom and Dad. Thank you for your unwavering support in allowing me to pursue any career that I wanted. Since I was little, you taught me what it was to have goals, to work hard, and to live my life with a sense purpose. Thank you for being my role models and teaching me the value of family. You are the strongest people that I know. I know that without both of you this, and all of my accomplishments, would not have been possible. I love you both so much.

Nothing in this world can take the place of persistence. –Calvin Coolidge

Thank you to Jason Allen – my partner, my better half, my biggest supporter. Thank you for all of your encouragement and support throughout my graduate education and being there to weather the storms with me. My parents instilled work ethic in me, but it's you who shows me each day how to be strong, overcome adversity, and persist. I could not have done this without you.

ACKNOWLEDGEMENTS

Colorless green ideas sleep furiously. –Noam Chomsky

Thank you to Paula Schwanenflugel for being the most amazing advisor. Thank you for being my sounding board through this process and taking the time to really invest in my work. I am so grateful for all of the opportunities you have provided for me. Without you, all of my most important experiences in graduate school would not have happened. When I first arrived at UGA, you gave me the opportunity to work on a project which later became this very manuscript. You then recommended me for an assistantship which later turned into a career. You have played an integral role in my support system throughout graduate school and there are not enough words to say how much I appreciate you.

Great spirits have often overcome violent opposition from mediocre minds. –Albert Einstein

To my committee members: Nancy Knapp, Stacey Neuharth-Pritchett, and Liang Chen. Thank you for providing me guidance through this whole process. Each of you brings a different perspective to the table and I know this manuscript has benefited from that.

All labor that uplifts humanity has dignity and importance and should be undertaken with painstaking excellence. –Martin Luther King Jr.

To my GCA/UGA family: Steve, Linda, Eliza, Michael, Ashton, Amanda, Cigdem, Bert, Hye-Jeong, and Donna. Thank you for holding me accountable for making progress and for giving me the day-to-day support to keep moving forward. Special thanks to Al for providing me the equipment and software needed to conduct my analyses. To Sara, Beryl, and Chris: it has

been the greatest pleasure sharing this experience with you. We will be forever bonded in this process. From carpooling to writing dates, thank you for making this journey fun.

Family is where life begins and love never ends. –Unknown

To my family: I am an only child, but I have never felt alone. Thank you to my aunts, uncles, and cousins for your life-long support and encouragement. I know that everything good in me is shaped by each of you. I love you all.

Family isn't always blood. It's people in your life who want you in theirs. The ones who accept you for who you are. The ones who would do anything to see you smile, and who love you no matter what. –Unknown

To my very best friends: Carin, Joanna, Marcus, and Nicole. You are the best family that I could have ever imagined choosing. Not just during graduate school, but for most of my life, you all have been a support system, a happy distraction, a phone call away. Through it all, you have stood by my side, and we will continue to stand together forever.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v
LIST OF TABLES ix
LIST OF FIGURES x
CHAPTER
1 INTRODUCTION AND LITERATURE REVIEW 1
Peabody Picture Vocabulary Test 6
Expressive Vocabulary Test 11
Comparing the PPVT and EVT 13
Lexical Diversity 19
Measures of Lexical Diversity 20
Lexical Diversity Measured by *D* 25
Kindergarten PAVEd for Success (K-PAVE) 33
Statement of Purpose 35
2 METHODS 38
Sample 38
Measures 45
Materials 48
Analytical Software 49
Procedures 49
3 RESULTS 53
Normality 53
Sensitivity to Sample Size 55
D's Relationship to Age 58
D's Relationship to Other Measures of Vocabulary 58
D's Relationship to Listening Comprehension 62
Differences between African American and Non-African American Children 63

4 DISCUSSION 66
The Validity of *D* Using Kane's Argument-Based Approach 66
D and Listening Comprehension 70
Differences between African American and Non-African American Children 71
Limitations of *D* 73
Future Research 73
5 CONCLUSION 75
REFERENCES 77
APPENDICES
A SIGNED RESTRICTED-USE IES DATA LICENSE 87
B SIGNED IES NOTARIZED AFFIDAVITS 96
C SIGNED IES SECURITY PLAN FORM 98
D LANGUAGE SAMPLE TRANSCRIBED IN CHAT SPECIFICATIONS 103
E FREQ COMMAND AND OUTPUT 104
F VOCD COMMAND AND OUTPUT 108

LIST OF TABLES

Page

Table 2.1: Characteristics of Students that Provided Language Samples as Percentages 40

Table 2.2: Steps for Determining Final Sample..... 42

Table 2.3: Comparison of Initial Sample and Final Analytic Sample 43

Table 2.4: Characteristics of Students in the Analytic Sample..... 44

Table 3.1: *D* for Different Parts of the Transcript Compared to Whole Transcript (*n* = 243)..... 56

Table 3.2: TTR for Different Parts of the Transcript Compared to Whole Transcript (*n* = 243). 56

Table 3.3: Correlations between *D* for Whole Transcript and Even- and Odd-numbered Words (*n* = 243)..... 57

Table 3.4: Descriptive Statistics of Language Variables (*n* = 244) 59

Table 3.5: Performance Differences between African American and Non-African American Children..... 64

LIST OF FIGURES

Page

Figure 3.1: Histogram with Normal Distribution Curve Overlay 54

Figure 3.2: Box-and-Whisker Plot with Outliers 55

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

The associations between vocabulary and reading comprehension, or the ability to understand what one has read, are well established in research (Hart & Risley, 1995), but questions about the most appropriate tool for measuring vocabulary in various populations remain unanswered. Various facets of vocabulary richness, including lexical diversity and use of unique words, have an important role in assessing language proficiency, monitoring progress, and testing theories of language development. Vocabulary is a fundamental factor in students' ability to comprehend what they read. At its most fundamental level, learning to read requires knowing the meanings of words in text (NICHD, 2000). The National Reading Panel, consistent with the *No Child Left Behind Act of 2001* (NCLB), asserts that vocabulary is one of five essential components in reading. Given the importance of vocabulary to reading comprehension, it is vitally important to establish a reliable and valid tool for measuring vocabulary acquisition for all populations. Hart and Risley (1995) assert that children from low SES families are exposed to much fewer words than children from middle-income families, and as a consequence, children from low-income families are likely to score lower on standardized vocabulary assessments. Given the disproportionate representation of minority children from families of low SES, and the variations of socialization norms, performance on vocabulary tests also seem to differ as a function of ethnic background (Stockman, 2000). This chapter will review the importance of vocabulary acquisition in young children, how vocabulary has been typically measured by standardized assessments, and the issues regarding these vocabulary assessments

for multicultural populations. Specifically, I will review the issue of cultural bias exhibited in standardized vocabulary assessments for at-risk, low-income, African American children. I will propose that the more recently developed *D* may serve as an unbiased measure of lexical diversity for these young children.

Research has shown that the importance of vocabulary extends beyond early childhood into young adulthood. In fact, vocabulary measured in preschool correlates with reading comprehension when measured in upper elementary school (Dickinson & Tabors, 2001). More specifically, receptive vocabulary and listening comprehension measured in kindergarten have been directly related to children's reading ability in third grade (Sénéchal & LeFevre, 2002). Similarly, vocabulary size measured in kindergarten is an effective predictor of reading comprehension in middle elementary years (Scarborough, 1998). Further, in an examination of the literacy progress in students who participated in Head Start, a federally-funded program to improve school readiness for young children from low-income families, Spira, Bracken, and Fischel (2005) found that kindergarten vocabulary skills were strongly associated with improvement in reading from first through fourth grade. Thirty percent of children in the sample demonstrated steady improvement through the end of fourth grade despite having deficient decoding skills in first grade. Further, Cunningham and Stanovich (1997) found that first grade vocabulary ability was a strong predictor of reading ability up to ten years later, accounting for over thirty percent of the variance in reading comprehension even when removing the effects of cognitive ability.

Vital to literacy development are the home experiences that children have. In fact, Hart and Risley (1995) found that vocabulary growth at age three was strongly correlated with family SES (r = .65). Research has suggested that the lower performance of many children from African

American homes and from families of lower SES results from a mismatch between the language experiences at home or in the community and the language typically expected and measured in schools.

A reader's background knowledge and experiences strongly influence reading comprehension. In reading, the familiarity of the vocabulary, concepts, and structural elements of a text contributes to successful comprehension (National Assessment Governing Board, 2012). Children of low SES are especially likely to struggle with reading. This struggle emerges early and continues into the elementary school years (Dickinson & Tabors, 2001; Hemphill & Tivnan, 2008). The experiences of children from low SES families differ from those of middle-income families, and these differences affect not only vocabulary development, but literacy skills in general. For example, children from low SES families are less likely to attend educationally-focused preschools. This is because both of the cost and availability of good preschools in the community (Fuller, Eggers-Piérola, Holloway, & Rambaud, 1996). Communities where children of low SES families live also have less access to print; that is, these communities may have fewer libraries and bookstores than middle income communities (Neuman & Celano, 2001). Furthermore, there could be a mismatch between the language patterns used in low SES families (i.e., in parent-child conversations) compared to those that would facilitate literary development as promoted in schools (Hoff, 2006). Taken together, the experiences that children who live in poverty bring to school not only mismatch the skills promoted in schools, but also do not match the experiences upon which standardized tests are based. This puts children from low-income families at risk for both reading difficulties and lower performance on standardized assessments (Hart & Risley, 1995; Stockman, 2000; Washington & Craig, 1999).

Improving the ability to read and comprehend text in at-risk children has been a high priority in education policy over the last two decades. One critical factor in reading achievement is adequate vocabulary knowledge (Goodson, Wolf, Bell, Turner, & Finney, 2010). Children who enter school with limited vocabulary knowledge grow more discrepant over time from their peers who have rich vocabulary knowledge (Baker, Simmons, & Kaméenui, 1998). Children from low SES families are exposed to one-third of the language that children from highly verbal families are exposed to (Hart & Risley, 1995). In a longitudinal study, Hart and Risley (2003) followed children from seven to nine months until they turned three years old. Researchers sampled children from 42 families with varying socioeconomic status – 13 families were upper SES, ten were middle SES, 13 were lower SES, and six were on welfare. Analyzing over 1,300 hours of child-parent interactions and extrapolating the observational data, Hart and Risley estimated that a child from a professional family would be exposed to almost 45 million words before entering school compared to 13 million for a child from a family on welfare. This vocabulary gap between struggling readers and proficient readers persists through the elementary school years (White, Graves, & Slater, 1990), with the discrepancy growing each year (Stanovich, 1986).

Given the negative outcomes associated with early language deficits, accurate screening and early identification are critical to guide intervention efforts (Neuman, 2009). Upon entering school, low SES children may know one-half or fewer of the word meanings known by children from middle-income families, and far fewer than children from high SES families (Hart & Risley, 1995). Therefore, children from low SES families likely enter school having less vocabulary knowledge than their middle SES counterparts. However, there is concern that standardized tests have been normed using a larger percentage of middle-income than low-

income children. Additionally, researchers have also raised concerns about the valid use of standardized language test in describing the language development of children from low-income, culturally and linguistically diverse backgrounds. For example, African American children have been shown to score lower than Caucasian children on standardized receptive vocabulary measures (Kresheck & Nicolosi, 1973). It is unclear whether the results from these tests are appropriate for identifying language delays in low SES minority children (Qi, Kaiser, Milan, & Hancock, 2006).

Standardized vocabulary tests, such as the Peabody Picture Vocabulary Test (PPVT) and the Expressive Vocabulary Test (EVT), are frequently used by speech-language pathologists or clinicians in school settings to screen for language impairments, diagnose reading difficulties, and measure response to vocabulary instruction or intervention. Although these standardized assessments are often quick and easy to administer, they have been criticized for being negatively biased when applied to multicultural populations. Stockman (2000) asserts that, because vocabulary tests focus on word meaning, and meaning is neither overt nor finite, vocabulary knowledge is more difficult to accurately assess compared to other aspects of overt language ability, such as, for example, knowledge of word surface structure. The words that a person acquires are influenced by cultural experiences. Even for known words, the boundaries of meaning are not fixed. Meanings of words can vary for people from different linguistic communities because of different cultural experiences. Similarly, a word can have multiple meanings even within the same linguistic community. Achieving an unbiased vocabulary assessment is difficult because of the different regions, age ranges, social classes, and ethnic groups being tested. It is important to examine available vocabulary tests and determine which measures are appropriate for multicultural populations. In the following sections, descriptions of

standardized vocabulary assessments the PPVT and EVT will be reviewed along with research that have investigated their validity when assessing culturally and linguistically diverse populations.

Peabody Picture Vocabulary Test

The Peabody Picture Vocabulary Test, Fourth Edition (PPVT-4; Dunn & Dunn, 2007) is a standardized assessment to measure receptive vocabulary. The PPVT-4 is one of the most commonly used instruments for assessing receptive language ability (Thomas-Tate, Washington, Craig, & Packard, 2006). The PPVT-4 asks the test taker to select the picture that best represents the meaning of a stimulus word presented orally by the examiner. For example, the examiner says the word 'bicycle' and the examinee points to one of four pictures presented in the booklet. The PPVT-4 requires no reading or writing and is quick and easy to administer individually. The PPVT has been used to measure response to vocabulary instruction, screen for verbal development, detect language impairments across age ranges, as well as a measure for verbal ability in large-scale federally funded preschool research projects, such as Early Reading First (U.S. Department of Education, 2009) and Even Start programs (Ricciuti, St. Pierre, Lee, Parsad, & Rimdzius, 2004). The fourth edition of the PPVT improves on the third edition by increasing the number of easy items at the beginning of the test to more accurately measure children who function at very low levels. According to the manual, measures of internal consistency range from .92 to .98 and test-retest reliability ranges from .91 to .94. Researchers believe that it is a relatively valid measure of children's verbal ability. However, evidence concerning the potential cultural bias of the PPVT is mixed.

Washington and Craig (1999) sought to determine the appropriate use of the PPVT-III with at-risk African American preschoolers. They sampled preschoolers from four state-

sponsored at-risk preschools in the Metropolitan Detroit area. Of the children in the preschools they sampled 59 at-risk African American students (*M* age = 51 months). Students were classified as at-risk based on "either income or other environmental factors . . . [such as] family density, single parent households, and/or significant family history (e.g., teenage parents)" (p. 76). Of 51 children whose income information was determined from school records, 14% had caregivers with less than a high school education, 71% had at least a high school diploma, and 18% had college degrees. The researchers found that their African American preschool sample scored lower (*M* score = 91, *SD* = 11) compared to the PPVT-III normative sample (*M* score = 100, *SD* = 15). However, despite scoring lower than the standard mean there was a range of performance in the at-risk sample and the difference in scores was not statistically significant. Therefore, Washington and Craig concluded that the PPVT-III was culturally unbiased for assessing receptive language skills of young African American children.

There is evidence that suggests that the PPVT is not valid for some groups of children, and therefore evaluation of programs might be compromised by the use of the PPVT–III, and the policy implications drawn from scores might be problematic for African American children. Campbell, Bell, and Keith (2001) tested the PPVT-III as a screening measure for intelligence and achievement in 416 African American kindergarteners (*M* age = 75.86 months; *SD* = 3.12 months) of low SES (*M* household income per year = $13,467). The researchers calculated correlations between the PPVT-III and the Kaufman Assessment Battery for Children (KABC; Kaufman & Kaufman, 1983), their measure for intelligence and achievement, to determine the concurrent validity of the PPVT-III. They found that correlations between PPVT-III, KABC's Mental Processing Composite IQ score, and KABC's Achievement Composite score were positive and moderately significant, ranging from .44 to .64. Similar to the results from

Washington and Craig (1999), researchers found that low SES African American students scored more than 1 *SD* lower than the mean (*M* score = 82.26, *SD* = 12.19) of the PPVT-III normative sample. The PPVT-III, on average, tended to underestimate achievement, as measured by the KABC in low SES African American children. The researchers concluded that, as an achievement screener, the PPVT-III did not achieve sensitivity and specificity values that met or exceeded .80, as suggested by Carran and Scott (1992).

Campbell and colleagues were also interested in examining how the PPVT-III compared to its previous versions and found that the PPVT-III and PPVT-R, the previous edition, were comparable, indicated by the similar relationships with KABC summary scores and like performance by low-income African American children. This particular finding contradicts the recent findings of Washington and Craig (1999) who concluded that PPVT-III was culturally unbiased for assessing receptive language skills of young African American children, in contrast to the PPVT-R, in which Washington and Craig (1992) concluded to be racially and economically biased toward low-income African American preschoolers and kindergarteners. The researchers determined that the distribution of PPVT-R scores for their subjects was significantly skewed toward the low tail of the standard normal distribution. Their results indicated that 91% of their subjects scored significantly below (*M* = 79.7, *SD* = 15.9) the standardized mean and there was evidence of item bias. These findings call into question the applicability of the PPVT to young low-income African American students.

To examine the use of the PPVT-III in a federally funded program that promotes school readiness, Champion, Hyter, McCabe, and Bland-Stewart (2003) administered the PPVT-III to forty-nine 3- to 5-year old African American children enrolled in a Head Start program. The researchers reported that nationally, Head Start serves families with annual incomes of less than

$9,000; therefore 100% of the families in their sample are considered impoverished. They found the children in their sample also performed significantly lower (*M* score = 86.84, *SD* = 10.96) than the mean for the normative sample. The researchers also reported that the scores from their sample were not normally distributed; they reported a leptokurtic distribution in their sample, indicating that there were few children with very high or very low test scores (80% of the sample clustered within 1 standard deviation of its own mean). This finding is also in contrast to the results from Washington and Craig (1999) who found that even though their sample of at-risk African American children scored below the standardized mean, the difference was not statistically significant and their subjects' scores did show a range in performance.

It's possible that the difference in results between these two studies can be attributed to the samples. In the study by Champion and colleagues, 100% of the children were from impoverished families, whereas less than half of the subjects in Washington and Craig (1999) were considered low-income. Further, students enrolled in the Head Start program may need more direct instruction to support language and literacy development than those not enrolled in the program. Head Start Family and Child Experiences Survey (FACES), published by the U.S. Department of Health and Human Services (U.S. DHHS, 2006), found that children from low-income families who were enrolled in a Head Start program were at a greater disadvantage as compared to other children, as evidenced by children's initial scores on standardized assessments of cognitive skills. According to the 2002 survey results, Head Start entrants had a mean standard score of 85.3 on the PPVT-III, and the typical Head Start child was at about the 16th percentile in receptive vocabulary skills. In all of the above studies investigating the validity of the PPVT it is difficult to determine whether the differences in performance between the subjects and the normative sample can be attributed to race, SES, or both, as all studies included African

American children of low SES. Some have posited that the differences in performance could also be attributed to the decontextualized tasks in the test.

The tasks of the PPVT require the test taker to listen to the examiner provide a verbal stimulus word and then point to the picture; this task may be too context removed, making it an unfamiliar task to this population of children. This may cause poorer performance when compared to the EVT. Children of low SES may not have the same ability to respond to decontextualized tests compared to children of higher SES, who may have experienced similar activities prior to entering school (Restrepo et al., 2006). Low SES families tend to have literary experiences as part of the contextual focus of conversations rather than as separate literacy activities. For example, Fagundes, Haynes, Haak, and Moran (1998) found that African American children performed better on a modified version of the Preschool Language Assessment Instrument (PLAI; Blank, Rose, & Berlin, 1978) when the test items were provided in the context of thematic activities (e.g. story, arts and crafts) than in the traditional administration of the PLAI where the stimulus is presented in black and white line drawings. It is important then to explore the same validity issues in other standardized measures of vocabulary. The PPVT-4, the latest edition, has been co-normed with the Expressive Vocabulary Test, Second Edition (EVT-2; Williams, 1997). This allows comparisons to be made directly between receptive and expressive vocabulary performance on these two assessments. These instruments are often paired in practice to obtain a more comprehensive picture of vocabulary knowledge (Thomas-Tate et al., 2006). The following section reviews studies examining the validity of the EVT for assessing expressive vocabulary skills in multicultural populations.

Expressive Vocabulary Test

The EVT is a standardized assessment that measures expressive vocabulary. The EVT is an individually administered, norm-referenced test designed for examinees of at least two and a half years old. At the beginning levels, the EVT asks test takers to label the word of the picture or a part of the body that the examiner points to. At the more advanced levels, the EVT asks test takers to provide one-word synonyms of labeled pictures. The test contains 190 items; the first 38 items are labeling items and the remaining 152 items are synonym items. For each item, the examiner first points to a picture or a part of the body and then asks a question. The test taker responds with a one-word answer that is a noun, verb, adjective, or adverb. The EVT manual suggests that the test can be used to screen for expressive language problems, measure word retrieval, understand reading difficulties, and monitor vocabulary growth. The EVT-2 has internal consistency (split-half) reliabilities of .94-.95 and has a test-retest reliability of .95 (Williams, 2007). Researchers believe that it is a relatively valid measure of children's verbal ability. However, evidence concerning the potential cultural bias of the EVT is mixed.

Thomas-Tate and colleagues (2006) sought to determine if the EVT (Williams, 1997) is a valid instrument for assessing expressive vocabulary skills in African American preschool and kindergarten students. Researchers sampled 165 African American preschool and kindergarten students (*M* age = 4.23 years, *SD* = .62) from two Michigan communities of different demographics – a mid-sized central city (with a large number of affluent and highly educated residents) and an urban-fringe city (with a large percentage of low-income families). Although the communities differed overall, the African American students from both were very similar; they were disproportionately poor and low-academic achievers. The mean EVT score for this sample was 96.42 (*SD* = 11.45) which was not significantly lower than the standardized mean,

and there was a wide range of scores suggesting sufficient performance spread. Additionally, although slightly lower, the mean score fell within 1 *SD* of the standardized mean. While it seems, based on this study, that African American students perform similarly on the EVT compared to the normative sample, it is difficult to make that conclusion because of their mixed income sample. Because the students were not further sub-grouped by SES or community it is possible that the scores of the students from the affluent neighborhood are masking the lower scores of the students from the urban-fringe city. If they had further sub-grouped their sample it is possible there would be performance differences between the two communities.

Thomas-Tate and colleagues also investigated the criterion validity of the EVT, examined by analyzing the relationship between EVT and PPVT-III scores which were also collected in the study. There was a strong, positive, and statistically significant correlation ($r = .66$, $p = .01$). These findings together led the researchers to conclude that the EVT was a valid measure of expressive vocabulary for African American students in the Midwest. However, further analyses revealed that EVT scores alone were not adequate in diagnosing language disorders. Twenty-nine students in the study failed the language screening battery (which included the PPVT, subtests of the KABC, a Wh-questions comprehension task, and a picture description tasks from which mean length of communication unit [MCLU] was calculated), but only four of these students had EVT scores lower than 1.5 *SD* below the mean (criteria often used to indicate a language disorder). Researchers concluded that the EVT is valid for measuring expressive vocabulary skills in this group of African American students, although used alone may not be able to identify language disorders.

Comparing the PPVT and EVT

Few studies have examined the results of the PPVT and EVT together in the same sample. In an effort to extend the findings from Thomas-Tate and colleagues (2006), McCabe and Champion (2010) examined whether the EVT (Williams, 1997) is culturally fair when sampling students of low SES from a southern state (rather than Michigan). McCabe and Champion sampled fifty-three African American children (18 kindergarteners, 21 second graders, and 14 fourth graders), all of whom participated in the free lunch programs at their schools. Each student completed the EVT, PPVT-III, and the Diagnostic Evaluation of Language Variation Screening Test (DELV-Screening Test; Seymore, Roeper, deVilliers, & deVilliers, 2009), used by clinicians to distinguish variations due to normal developmental language changes or cultural patterns of language differences from true markers of language disorders or delay. Results indicated that their sample's mean score on both the EVT (M = 90.58, SD = 13.18) and the PPVT (M = 91.55, SD = 13.47) were significantly lower than that of the normative samples. These results do not support those found by Thomas-Tate and colleagues (2006), who did not find a statistically significant difference in EVT scores between their sample and the normative sample. The sample size and the fact that children were not separated into different age groups are limitations of this study. The range of ages in this study spans at least four grade levels. Even so, the findings from this study are consistent with the results found by other researchers (Qi et al., 2006; Restrepo et al., 2006). Further, this provides some indication to regional differences between the South and the Midwest.

The second purpose of the McCabe and Champion study was to compare PPVT-III and EVT scores to determine if the EVT was found to be an easier test than the PPVT-III for low-income African American children. The results did not support their hypothesis. These findings

contradict other researchers (Restrepo et al., 2006) who did find that there was significant difference in scores from the EVT and PPVT-III in their sample. This difference could be attributed to their different samples. Rather than sampling from the same age group, this study sampled 53 students of varying ages (in kindergarten, 2nd grade, and 4th grade). Results in this study were not reported by grade; that is, results were only reported for the entire sample and therefore reasons for the discrepancy are inconclusive.

Few researchers have examined the fairness of the PPVT and EVT using a comparative sample of both African American and European American children in the same study. Qi and colleagues (2006) examined the validity of the PPVT-III by comparing the performance of African American preschoolers (n = 482) and European American preschoolers (n = 42) from low-income families eligible for attending Head Start or low-income child care programs. They hypothesized that there would be no difference in performance. Further, to assess the convergent validity of the PPVT-III, correlations between the PPVT-III, EVT, and Preschool Language Scale-3 (PLS-3) were computed. Hierarchical multiple regressions were run to test whether the PPVT-III scores were related to selected SES factors. Researchers found that the mean score for their African American sample (M = 77.88, SD = 13.19) was approximately 1.5 SD below that of the standardized mean and corresponded to approximately the seventh percentile of the normative sample. When compared to the European American students (M = 81.90, SD = 16.00), African American students performed slightly lower, however an independent samples t-test on the mean scores for each group revealed that the difference was not statistically significant. This finding is consistent with the results from other studies (Campbell et al., 2001; Washington & Craig, 1999). Given the small sample of European American students and the discrepancy in group sizes, the authors warned that the results should be interpreted cautiously. These findings

suggest that group differences may be attributed to SES rather than racial bias in the PPVT. When PPVT-III scores were correlated with the EVT and the PLS-3, results indicated strong correlations (r = .63 and r = .56, respectively). Additionally, students who performed lower on the PPVT (< 70) had significantly lower scores on the other two measures as well. The two groups of students performed similarly on the EVT, but because of the uneven sample sizes, and the strong correlations between EVT and PPVT scores, results should be interpreted with caution.

Restrepo and colleagues (2006) investigated the validity of the PPVT-III and the EVT for assessing vocabulary in preschool children. The researchers sampled two hundred and ten 4-year-olds from a lottery-funded public preschool program in Georgia with 30% of the sample receiving free or reduced lunch; of the total sample, 57.6% were African American and 42.4% were European American. Researchers were also interested in whether the PPVT-III and EVT were unbiased for children with varying maternal education levels, sometimes used as an indicator to determine SES. Ten percent of the total sample had mothers with less than a high school education (67% African American [AA] and 33% European American [EUA]). Fifty-seven percent had mothers who completed high school or a GED (63% AA and 37% EAU), eleven percent had mothers who completed at least some college (54% AA and 46% EAU), and twenty-two percent who had mothers who completed college (41% AA and 59% EAU). Fifty percent of the sample was females and all children reported English as their first language.

Restrepo and colleagues hypothesized children whose mothers have less than a high school education will score significantly lower than children whose mothers have high school or greater education levels. Further, they hypothesized that African American children would score similarly to their European American peers and that differences in performance would be mostly

due to the mother's education level. Results indicated that African American children (*M* = 84.21, *SD* = 12.79 for the PPVT-III and *M* = 93.83, *SD* = 11.78 for the EVT) scored lower than European American children (*M* = 101.84, *SD* = 11.49 for the PPVT-III and *M* = 102.18, *SD* = 11.49 for the EVT) on both tests. Children whose mothers did not complete high school (*M* = 77.95, *SD* = 13.02 for the PPVT-III and *M* = 90.57, *SD* = 7.72 for the EVT) tended to score lower on both measures than children whose mothers completed high school or higher; however, this effect was larger for the PPVT–III than the EVT . They found that 34.8% of the sample scored significantly lower on the PPVT-III than the EVT but only 11% vice versa. Further, African American students performed better on the EVT than on the PPVT-III, but this was not the case for European American students who scored nearly identically on both tests. Researchers concluded that the EVT was a better indicator of vocabulary skills than the PPVT-III, and it was less likely to result in the placement of African American children in low achieving groups. This is consistent with the conclusions by Qi and colleagues (2006). Similarly children whose mothers had low education levels also performed better on the EVT than on the PPVT. Discrepancies between scores on the PPVT and EVT grew larger as mother's education level declined.

To specifically investigate if there are items on the PPVT-III that exhibit differential item functioning (DIF), Webb, Cohen, and Schwanenflugel (2008) used latent class analysis (LCA) to examine if any individual items exhibited racial bias. Using standard DIF analyses, Webb et al. found that only one item of twenty-four was detected as exhibiting DIF and it favored the African American students in their sample. Using LCA, the authors selected a two-class model resulting in lower vocabulary scores in Class 1 and higher vocabulary scores in Class 2. Class 1 was comprised of 74.8% African American students (*n* = 83) and Class 2 was comprised of

79.5% European American students (n = 58). Researchers found that the effect of latent class (or ability) is much larger than ethnicity, with Class 1 showing lower vocabulary ability on both the PPVT and EVT than Class 2. Further, mixed ANOVA analyses revealed that only students in Class 1 performed significantly lower on the PPVT than on the EVT. These findings are surprising because theory on how words are acquired would expect that receptive vocabularies are larger than expressive vocabularies. The authors posited that reasons for a disproportionate amount of African American students in Class 1 include differences in familiarity with the task of the PPVT, differences in meaning for any given picture, or a dialect mismatch between examiner and examinee.

There may be "no ultimate solution to the problem of culture-fair measurement" (Sternberg, 1988, p. 82), thus it is important to consider methods with the potential to reduce bias in language assessment. The PPVT-R was an attempt to reduce cultural bias in a standardized assessment, but it was determined to be inappropriate for the use of most young children from different cultural and linguistic communities, including African American children (Washington & Craig, 1992). The combined results from the aforementioned studies that compare the performance of African American children of low SES on the PPVT and the EVT (McCabe & Champion, 2010; Qi et al., 2006, Restrepo et al., 2006; Thomas-Tate et al., 2006) favor the EVT as a more culturally fair assessment than the PPVT (mean scores for EVT were higher in each study except for McCabe and Champion but the difference was not statistically significant). However, because of the mixed results, seemingly lower performance by African American children of low SES, alternative nonstandardized measures should be explored. This has been investigated to some extent.

Campbell, Dollaghan, Needleman, and Janosky (1997) identified processing-dependent language measures that reduce the emphasis on a subject's prior vocabulary knowledge and experience. One of their measures, Nonword Repetition Task (NRT) was determined to be a culturally fair procedure for assessing minority children that does not penalize for lack of world knowledge. It is designed to evaluate phonological working memory storage by asking examinees to repeat series of nonwords. There were no differences in performance on the NRT for White and minority children unlike performance on the Oral Language Scale, a knowledge-dependent language measure, in which the minority group scored significantly lower. Deficits in nonword repetition could reflect fundamental deficits in phonological working memory storage that are associated with lexical and language development (e.g., Gathercole & Baddeley, 1993). The NRT has been shown to be effective as a language screening instrument in African American children in preschool and kindergarten when considered with other nonstandardized measures (Washington & Craig, 2004).

Another approach to screen for vocabulary issues in language development is to engage the child using a familiar medium like conversation, record their speech, and then analyze for the diversity of vocabulary used in the output. In this study, I will evaluate the potential for lexical diversity to serve as a relatively unbiased measure of vocabulary skill for groups not typically served well by standardized vocabulary measures (i.e., children from low income families, African-American children, children from the Southern U.S.). In the subsequent sections, I will review measures of lexical diversity, studies investigating the validity and reliability of *D*, and studies that have used *D* as the measure for lexical diversity in various populations.

Lexical Diversity

Lexical diversity, or lexical variation, is measured by comparing the number of unique words with the total number of words in a language or writing sample (Malvern, Richards, Chipere, & Durán, 2004). Lexical diversity has been used as a variable in a wide range of research topics including assessments of first and second language ability, autism, language delay, and language impairment (Durán, Malvern, Richards, & Chipere, 2004; Jacobson & Walden, 2013; Klee, Stokes, Wong, Fletcher, & Gavin, 2004; Malvern & Richards, 2002; Malvern et al., 2004; Owen & Leonard, 2002; Silverman & Ratner, 2002). Lexical diversity has also been a dependent variable in intervention studies focusing on vocabulary development in low performing prekindergarten children (Ruston & Schwanenflugel, 2010). Lexical diversity does not refer to all the words a child can comprehend, but rather gauges lexical ability that occurs during (relatively) spontaneous speech. Unlike standardized vocabulary tests, which ask test takers to provide definitions and synonyms, or label or identify a stimulus/target word, lexical diversity can estimate someone's active vocabulary that they may know well enough to use in everyday speech or writing (Malvern & Richards, 2002). An advantage to using lexical diversity as an indicator of vocabulary knowledge is that it removes background knowledge and cultural differences that may bias scores on standardized assessments. Children can use as many or as few words that they choose in their language productions. Lexical diversity measured in language samples allows the measure of vocabulary ability using a familiar medium (e.g., conversation) without necessarily biasing against different cultural upbringings. There are various measures of lexical diversity that have been used, but the recent development of *D,* a measure based on mathematically modeling how new words are introduced into language samples, overcomes some of the shortcomings of other measures previously used. In the

following section, various measures of lexical diversity will be reviewed and an argument for *D* as the most valuable measure to assess lexical diversity will be provided.

Measures of Lexical Diversity

Number of different words (NDW). Calculating for number of different words (NDW) is the simplest measure of lexical diversity. Studies by Miller (1991) and Klee (1992) have shown that NDW is strongly correlated with age, and therefore an indicator of development. Additionally, NDW differentiated between normally developing children and children with specific language impairments (SLI). Although more recently it was found that NDW did not differentiate between typically developing children and children with bilingual language impairments (BLI; Jacobson & Walden, 2013). Even so, there are flaws to the procedures used to collect NDW. First, NDW is limited when comparing the lexical diversity of different transcripts because it is dependent on the size of the language sample. Given two transcripts, it is likely that the longer transcript will create more unique words than a shorter transcript. Various studies (e.g., Klee, 1992; Thordardottir & Ellis Weismer, 2001) have tried to standardize the sample by controlling for length as measured by utterances (continuous piece of speech beginning and ending with a clear pause), time (i.e., speech in one minute), and number of words (i.e., out of 300 words). Each method has its flaws – even when controlling for number of utterances, number of words per utterance will vary. Additionally, when controlling for time, researchers found that number of words per minute, for example, vary by age and language ability. When controlling for number of words, the percent of transcript cut (i.e., percent not analyzed) varies by transcript and a transcript with a higher percent cut will not be as representative of speech as a shorter transcript (i.e., a 600 word transcript cut down to 300 words is not as representative as a 350 word transcript cut down to 300 words). Some studies have also tried random sampling of

words to calculate lexical diversity, where 25 words, for example, are randomly sampled in the transcript with or without replacement, that is, the same word can be selected again and again. The varying methods of standardization produce unreliable results and the results will vary depending on which method is employed to calculate NDW as a measure of lexical diversity. Although NDW maybe be an effective measure for discriminating typically developing children from children with language disorders, this does not provide a measure of lexical diversity since it confuses volubility with lexical skills (Owen & Leonard, 2002).

Type-token ratio (TTR). Originally proposed by Johnson (1944) type-token ratio has been one of the most influential measures used to quantify lexical diversity. *Type* refers to the number of unique words used and *token* refers to the total numbers words used in a transcript. Therefore, t*ype-token ratio* refers the proportion of words that are unique compared to the total number of words expressed. A higher ratio represents greater diversity in a transcript. Because TTR is a proportion, there is an assumption that it is not dependent on sample size, and therefore a more robust measure than NDW. However, higher values are obtained from shorter samples and similarly, lower values from larger samples. In Malvern et al. (2004), the authors reported that when a Prince of Wale's speech was analyzed, the TTR for the first 365 words in the speech was 0.60, but only 0.36 when the entire speech was analyzed (834 types from 2350 tokens). Similar to NDW, a standardization method needs to be employed. Studies (Broen, 1972) have standardized by time (i.e., five-minute segments of recording), or by analyzing the middle 50 or 100 consecutive utterances in a transcript. A major limitation of using TTR as a measure for lexical diversity is that in order for results to be comparable across research studies, there needs to be an agreement on the number of tokens required to estimate the TTR or an agreed upon

standardization method. Further, if the topic of speech is focused, then it is likely that the speaker will reuse content words related to that topic thus decreasing TTR.

Mean segmental type-token ratio (MSTTR). The most successful method of standardizing TTR is to base the calculation on an average of a number of subsamples of fixed token size. A transcript can be segmented based on the specified number of fixed tokens (e.g., 100 tokens per segment). For each segment the TTR is calculated and then averaged for all segments. Taking subsamples makes use of the data available by taking the mean of the maximum number of segments present in the transcript. The segments should be small enough so that multiple segments can be analyzed for the shortest transcript. Because most of the transcript is being analyzed in multiple segments it increases the reliability. Even so, there are issues with using MSTTR as a measure of lexical diversity. First, different size segments are not comparable; that is, a transcript segmented into 100 word segments (MSTTR100) cannot be compared to a transcript segmented in 30 word segments (MSTTR30). That is because TTR falls as tokens increase (Malvern et al., 2004). Second, segments don't take into account repetition outside of each segment. Third, there is still some loss in data because not all transcripts will segment into equal parts. Fourth, the relationship between types and tokens is dynamic; MSTTR represents a single point on a curve representing the way TTR decreases with an increase in token size (Malvern & Richards, 2002).

Moving-average type-token ratio (MATTR). Moving-average type-token ratio (Covington, 2007; Covington & McFall, 2010) measures lexical diversity by calculating TTRs for successive nonoverlapping segments of a sample. Once a window length is selected (e.g., x = 500 tokens) the TTR for tokens 1 to 500, for example, is estimated. Next, the TTR is estimated for words 2 to 501 (x + 1), then 3 to 502 (x + 2) and so on for the entire sample. The final

MATTR value is the average of the estimated TTRs. Thus, MATTR yields a value for every point in the text, while MSTTR is only a stepwise approximation to this. MATTR improves upon MSTTR by tracking the changes within text. Covington and McFall (2010) recommend that the window length should be smaller than the smallest text being analyzed, but large enough to provide a meaningful measure of style. Further, the authors also suggest that if the goal of analyses is to determine language/writing style then the recommended window length is 500; however, a much larger window (e.g., 10,000 words) is more appropriate if the goal is to determine the size of the speaker's vocabulary, such as in the case of lexical diversity.

Recently, Fergadiotis, Wright, and West (2013) evaluated the validity of four measures of lexical diversity in people with aphasia using confirmatory factor analysis to determine whether all computational tools measure the same latent variable and to what extent. Researchers analyzed 101 language samples from adults (*M* age = 63.09 years, *SD* = 11.32 years) obtained from a shared database of recordings of discourse from people with aphasia. Researchers found that all four measures, measure of textual lexical diversity (MTLD), MATTR, *D*, and TTR, reflect the same latent variable and were strong indicators of it. Moving-average type-token ratio was a strong indicator of lexical diversity that did not show evidence of systematic length effects and had a small residual variance, unlike TTR and *D*. Researchers concluded that these findings favored MATTR as a valid measure of lexical diversity that was not influenced by sample length. Unfortunately, a large window (e.g., 10,000 words) is needed to determine the size of a speaker's vocabulary and such a large window is not typically obtained studies with young children.

D. *D* was developed by Malvern and Richards (2002) as a quantitative measure of lexical diversity and is particularly useful in research with child language samples because it adjusts for

the size of transcript. Overcoming the disadvantages of other measures of lexical diversity, the authors assert that *D* is not dependent on sample size, allowing valid comparisons between language samples of varying quantities of linguistic data. *D* randomly samples the whole transcript taking into account repetitions and uses all of the data. Each point on the curve is calculated from averaging the TTRs of 100 trials on subsamples consisting of the number of tokens for that point, drawn at random from throughout the transcript. The measure *D*, it is argued, is the most reliable measure of lexical diversity and is particularly useful for measuring short transcripts such as those produced by young children.

The approach for calculating *D* is based on the probability of introducing new vocabulary into progressively longer language samples. It uses repeated calculations of TTR over a range of tokens (35 to 50) to show how TTR changes in relation to sample size. For example, type-token ratio is calculated one hundred times on 35 tokens selected at random without replacement within that sample of trial. That is, once a word in the sample is selected for a trial of 35 tokens it will not be selected again in the subsequent trials of 35 tokens. This is done again for 36 tokens and so forth until 50 tokens. Sixteen total points are plotted creating the empirical curve. The plotted points are then compared to a mathematical model curve of TTR, and that comparison (best fit) yields the value for *D*. Because *D* is calculated from a random selection of tokens within the sample, *D* values will slightly vary each time you calculate it. Therefore, *D* is calculated three times and then average of all three *D* values is the final "optimized" value used for reporting. *D* therefore provides a mean for measuring lexical diversity that "is not a function of the number of words in a sample, uses all the data available, and is more informative [than TTR] because it represents how the TTR varies over a range of token size" (MacWhinney, 2000, p. 128). Although *D* is not measured in units that are as transparent as NDW, higher *D* values

indicate greater lexical diversity (Malvern & Richards, 2002). Data from various studies (including children and adults, English and foreign speakers, normally developing children and those with specific language impairments [SLI]) confirm the methodological advantages of *D* as a better indicator of language developmental trend than other measures of lexical diversity.

Lexical Diversity Measured by *D*

Vocabulary diversity is determined by range, richness, and creativity of vocabulary (Ruston & Schwanenflugel, 2010) and can be quantified as *D*. Developed by Malvern and Richards (2002), it has been reported that *D* scores have been found to range from 1.48 at 18 months, to 119.20 for a sample of academic writing (Durán et al., 2004). Because of the mathematical modelling procedure used to calculate *D*, lexical diversity values are not a function of text length as with other measures (e.g., NDW and TTR). In this way, the developers assert that *D* overcomes the disadvantages of other measures. This is not to say that there is zero relationship between *D* scores and number of words, but that the value of *D* should not be a *function* of sample size. Lexical diversity is influenced by a number of factors such as the number of topic changes and language ability; advanced language users say more (i.e., may produce more words) and use more diverse language and therefore yield higher *D* scores. The reliability and validity of *D* has been investigated by the developers.

Malvern and colleagues (2004) analyzed transcripts of 38 children in the 32-month directory of the New England Corpus in the Child Language Data Exchange System (CHILDES) database. The children were all from English-speaking families and their ages ranged from 27 to 33 months (*M* age = 30.3 months). Because the *vocd* program, used to calculate *D*, uses random sampling of the transcript to obtain a *D* value, it was important to determine that there were no significant differences in *D* scores for the exact transcript if analyzed multiple times. The

stability of *D* was confirmed when a *t*-test revealed no significant differences ($t = 0.407$, $df = 37$, $p = 0.686$). Furthermore, the rank order of the children was maintained perfectly ($R = 1.00$). To investigate the internal consistency of *D* scores, developers used the split-half method, correlating *D* scores for even- and odd-numbered words and found a Pearson correlation of 0.763 ($df = 36$; $p < .01$; also reported in McKee, Malvern, & Richards, 2000). Further, there were no significant differences when comparing the *D* scores of even- and odd-numbered words to the total transcript when using random sampling without replacement (which has since become the default method of sampling in the *vocd* program). These results provide evidence that *D* is not dependent on sample size. The developers also investigated the criterion-related validity of *D* by correlating *D* with mean length of utterance (MLU) and found significant and moderately strong correlations, indicating *D*'s relationship with language development. Differences between groups of working-class and middle-class children provide evidence of discriminant validity. Further, developers found significant correlations between *D* scores and scores on the production measure of the Communicative Development Inventory (CDI) at both 14 and 20 months of age. They did not, however, find a significant correlation between *D* scores and the scores on the receptive measure of the CDI. Additionally, despite having a relative narrow range of ages, *D* did significantly correlate with age. The developers continued to investigate *D*'s relationship with age.

Durán and colleagues (2004) analyzed language samples originally collected between 1973 and 1978 from school children in Bristol, England (referred to as the Bristol Cohort, also reported in Malvern et al., 2004). Researchers analyzed transcripts of 32 children, with transcripts recorded at three-monthly intervals from 18 months to 42 months (9 total transcripts per children). For 15 of the children, transcripts were further collected at 5 years old. Researchers

found that the mean *D* score for the 5-year-olds was 64.02 (*SD* = 8.46, min = 50.83, max = 83.30). Further, researchers found that *D* is significantly correlated with age as a whole and for individuals in 87% of the cases. As children got older, the mean *D* scores steadily increased from 14.80 for 18 month children to 64.02 for 5-year-olds. These findings demonstrate a continuous trend for lexical diversity for early language development. At 39 months, children were also administered the English Picture Vocabulary Test (EPVT; Brimer & Dunn, 1963), a standardized test of receptive vocabulary. Researchers found that *D* measured at 39 months did not significantly correlate with scores on the EPVT. Researchers concluded that "*D* measures how diversely vocabulary resources are deployed and is likely to be related more to productive language than to assessments of receptive vocabulary like the EPVT" (p. 233).

Malvern and Richards (2002) investigated the convergent and discriminant validity of *D* by examining correlations between *D* and other student measures of language ability. In this study, lexical diversity is measured using samples from speakers who are learning a second language. Audio tapes were analyzed of British secondary school students taking a national oral exam for French at the age of 16. For the 27 students in their sample, researchers found that students' *D* scores correlated with MSTTR-30 (MSTTR for 30 tokens), another measure of vocabulary diversity, but did not correlate with the General Certificate of Secondary Education, which assesses general language ability. Further, they found no correlation between *D* and TTR ($\rho = .20$) nor *D* and total number of words, token ($\rho = .18$), but *D* was significantly correlated with the number of different words used, type ($\rho = .35$).

D has also been used in Hong Kong to measure the conversational language abilities of young Cantonese-speaking children. Klee and colleagues (2004) sampled seventy-four preschool and kindergarten students (age range 27-68 months) with and without specific language

impairments (SLI). In their first study of typically developing children, Klee et al. found that the *D* scores of their Cantonese-speaking sample were similar to the *D* scores of the children in the Bristol Cohort whose ages ranged from 18-42 months and 60 months (Durán et al., 2004). Additionally, because of the range of ages in their sample, researchers were also able to determine that *D* scores and age had a linear relationship. *D* scores ranged from 37.22 to 58.03 for students with the mean ages of 28.9 months and 47.6 months, respectively. The correlation between age and *D* scores was .71. This finding is consistent with their second study in which researchers compared children with SLI to typically developing children of the same age and to children of similar language ability as measured by a standardized test of language comprehension (RDLS-C). Typically developing children had significantly higher *D* scores than children with SLI of the same age and older children with SLI had similar *D* scores to younger typically developing children. The significant differences between groups indicate that *D* is sensitive to developmental trends.

Similarly, Owen and Leonard (2002) hypothesized that children with SLI may have lower lexical diversity than typically developing children of the same age, and that the lexical diversity of children with SLI may resemble that of younger typically developing children. Language samples from 144 children ranging from 2;2 to 7;3 years old (years; months) were analyzed, of which 53 children were diagnosed with a SLI (as determined by their scores on several expressive vocabulary measures) and 91 were typically developing children. All available transcripts were first cut to 100 utterances and each group of children was further sub-grouped by age (older and younger with at least 6 months separating the two groups). *D* and mean length of utterance in words (MLUw) were calculated. Consistent with previous findings (Durán et al., 2004; Klee et al., 2004; Malvern et al., 2004), the older group had higher *D* values

than the younger group. Similar to the results of Klee et al.'s Cantonese speaking sample, the researchers found that children with SLI scored significantly lower than typically developing children of the same age. When the children were sub-grouped by MLUw (higher and lower), they found higher *D* values for children with higher MLUws. Together, these findings support *D* being sensitive to developmental differences.

In a secondary analysis, researchers cut transcripts to 250 words using the same aforementioned groups. Their findings were consistent with the above results where older children had higher *D* values than younger children; however when comparing children with SLI and typically developing children, differences were no longer present between the two groups using the 250 word samples as they were in the 100 utterance samples. Further, researchers found that *D* values for the 100 utterance samples were larger overall than the 250 word samples. Concerned that *D* may be influenced by sample size, Owen and Leonard compared the *D* values when transcripts were cut to 250 words and cut to 500 words. Researchers found that for both children with SLI and typically developing children, *D* values were significantly different for the two sample sizes, although the effect sizes (Cohen, 1988) were small to medium (d = .35 and d = .43, for 250 words and 500 words, respectively). Researchers discussed that *D* may not be independent of language sample size. These findings are not surprising as the developers discussed a strong possibility of *D* being related to the number of tokens in a sample. Again, children with higher lexical diversity are likely to produce more words. Because the language samples in this study were limited to a certain amount of tokens, it is impossible to tell what was left out of analyses (i.e., what was cut). The results from this study can't be compared to the results of Malvern et al. (2004) because of the differences in the method to which the transcripts were analyzed (split-half method using odd- and even-numbered words).

D has also been used as to compare typically developing students to students with bilingual language impairment (BLI; Jacobson & Walden, 2013). Students with BLI have been found to perform lower on behavioral language measures relative to other bilingual children who have similar amounts of exposure to the language in question. They exhibit slower rates of vocabulary acquisition and higher rates of grammatical errors. *D* scores were generated from both English and Spanish language samples for all bilingual students. For English, in this case the second language being learned, there was a marginal difference ($p = .053$) between the language samples for typically developing children and children with BLI. Such difference was not exhibited in the Spanish language samples, in this case the native language. The findings differ a bit from the previous studies reviewed because *D* failed to obtain an effect for Spanish vocabulary, but this finding could be confounded to the extent that the sample was receiving English input at school.

Few studies have compared *D* scores to scores on standardized vocabulary measures. Silverman and Ratner (2002) compared *D* to TTR in a study of young children who stuttered ($n = 15$, *M* age = 35 months) compared to typically developing children of the same age ($n = 15$, *M* age = 35.67 months) and found that *D* distinguished between the two groups of children, but TTR did not. Not surprisingly, the correlation between *D* and TTR was extremely low ($r = .02$, $p = .92$) when the full language sample was analyzed, but increased when limiting the language sample to the middle 50 utterances ($r = .34$, $p = .07$). Further, *D* scores correlated ($r = .48$, $p = .01$) with performance on the expressive vocabulary test, Expressive One-Word Picture Vocabulary Test – Revised (EOWPVT-R; Gardner, 1990) whereas TTR did not. Similar to previous findings, *D* scores did not correlate with the receptive vocabulary measure, PPVT-R. These finding are consistent with the results from Malvern et al. (2004). The significant

correlation between *D* and a standardized expressive vocabulary measure provides some evidence of concurrent validity for *D* as a measure of lexical diversity, although results are limited because of the small sample size in this study. Additionally, children who stutter had significantly lower *D* scores than their fluent peers. This finding aligns with the research on children with SLI (Klee et al., 2004; Owen & Leonard, 2002) and children with BLI (Jacobson & Walden, 2013), which concluded *D*'s ability to detect developmental differences. The authors also tested *D*'s stability with varying sample lengths. Unlike Malvern et al. (2004), who performed split-half analyses on odd- and even-numbered words, Silverman and Ratner analyzed odd- and even-utterances compared to the whole sample. Results support the assertion that *D* is independent of sample size; there were no significant differences between odd- and even-utterances nor were there differences when each was compared to the whole sample.

Ruston and Schwanenflugel (2010) took further steps by using *D* as a dependent variable examining the effects of a 500-minute intervention on vocabulary skills, as measured by EVT and *D* scores, in preschool students (*N* = 73; *M* age = 4;4 [years; months], *SD* age = 3 months). The intervention was short-term and intensive, focusing on providing conversation that highlighted the use of rare words and open-ended questions that encouraged child talk. The preschools served a range of income levels and the children in the sample had average or below-average expressive vocabulary scores. The researchers hypothesized that although the intervention might benefit all students, it would be particularly beneficial for students with low initial vocabulary skills. Through repeated-measures analysis of variance (ANOVAs), the researchers found that the intervention group showed greater growth on the EVT than the control group, but the intervention did not significantly impact *D* scores. Further analyses were carried out to determine if the intervention affected students with low vocabulary pretest scores

differently than students with typically developing vocabulary. Researchers defined low vocabulary as children having pretest scores within the bottom third of normal distribution, or ≤ ½ *SD* below the normative mean. Given that the data were distributed normally (kurtosis and skewness values were acceptable), the researchers suggested this was an acceptable procedure for distinguishing groups. Therefore, students were considered to have low vocabulary skills if they scored between 73 and 93 on the EVT (43% of the sample) and between 22 and 45 on *D* (32% of the sample). Through two separate Group x Time partial repeated-measures ANOVA, analyses revealed that students who began the intervention with lower vocabularies benefited from the intervention but students who began with average or good vocabulary skills did not benefit. This was true for both measures, EVT and *D*. Although the results from this study suggest positive impact of a short-term intensive intervention on students with lower vocabulary skills, there are several limitations. First, the intervention was carried out by the researchers rather than the children's regular classroom teacher. Teachers may have difficulty fostering the same growth in young children's language skills as researchers when implementing interventions (Mol, Bus, & deJong, 2009). A meta-analysis on the effects of vocabulary intervention on young children's vocabulary acquisition revealed that larger effect sizes occurred when the experimenter conducted the treatment (Marulis & Neuman, 2010). Second, the intervention was not carried out in the classroom but rather in a separate quiet area. Together, these methods limit the generalizability of the results from this intervention if employed as regular classroom practice.

A considerable amount of research has investigated the appropriateness of standardized vocabulary measures on multicultural populations and results from these studies have been mixed. Standardized vocabulary assessments, like the PPVT, although easy to administer, have

been shown to contain bias specifically toward at-risk low-income African American children. Although other standardized measures, like the EVT, have been shown to be less biased, concerns still remain about the background knowledge and experiences valued by these standardized measures. *D* is an alternative approach to measuring expressive language. Unlike the standardized assessments reviewed, *D* is not assessing vocabulary by sampling commonly known words and meanings, but rather, relies on a sample of speech exhibiting the expressive vocabulary that the child actively uses. *D* has already been shown to be sensitive to developmental trends, and research with children with SLI and BLI has supported its use as a diagnostic tool for language impairments. Relatively little is known about the applicability of *D* to at-risk low-income African American children, who have traditionally scored lower on standardized measures compared to the norming sample and to European American children in the same age groups.

Kindergarten PAVEd for Success (K-PAVE)

In 2008, a kindergarten adaptation of a previously successful preschool vocabulary instructional program, PAVEd for Success (Hamilton & Schwanenflugel, 2011; Schwanenflugel, Hamilton, Neuharth-Pritchett, Restrepo, Bradley, & Webb, 2010) was implemented in public school kindergarten classrooms in the Mississippi delta region. The goal of the kindergarten vocabulary intervention was to examine the effectiveness of this adapted program (K-PAVE), as a supplement to literacy instruction, for improving vocabulary skills of children in this region. The Mississippi Delta region is plagued by high poverty rates, low student achievement, and its public schools draw from predominantly rural and African American communities. The study was conducted during the 2008-2009 school year with the initial training for teachers taking

place in the fall 2008 (see Goodson et al., 2010 for the detailed report). The current study uses assessment data collected from this regional study.

The K-PAVE program has three key components: Explicit Vocabulary Instruction (called "New VEhicles"), Interactive Book Reading (called "CAR Talk"), and Adult-Child Conversations (called "Building Bridges"). New VEhicles (Vocabulary Enhancement) involves explicit instruction of target vocabulary words using word-learning strategies and exposure to new vocabulary words in storybooks through repeated readings. The goal of the New VEhicles component is to provide explicit age-appropriate procedures for introducing new vocabulary, and to provide hands-on activities to extend student understanding of the target vocabulary words. CAR Talk involves teacher engagement of children during story reading by asking questions that promote comprehension and facilitate oral language skills. To implement CAR Talk in classrooms, teachers read each book more than once during the week, introduce books by taking a "book walk", and read the books interactively by using prompts based around the acronym CAR (Competence, Abstract, and Relate). Students can benefit from multiple readings of the same book because they often miss vocabulary words or key concepts during the first reading. Further, using prompts/asking questions engages students in the book and uses the book as a vehicle for conversation. The Building Bridges component involves individual or small group conversations with students to provide an opportunity for the teachers to use new vocabulary and for the students to increase their productive use of new vocabulary and their oral language skills. The program recommends 15 minutes per small group per week, split up into three 5-minute sessions. Together these components promote student vocabulary through multiple pathways.

The researchers of the K-PAVE study (Goodson et al., 2010) were interested in determining the impact of the K-PAVE program on student's expressive vocabulary, listening

comprehension, and academic knowledge in kindergarten. They were also interested in the impacts of K-PAVE on kindergarten instructional practices and if K-PAVE teachers spent less time on nonvocabulary literacy teaching practices (e.g, print, phonological awareness, alphabet knowledge) compared to teachers in the control group. Researchers found that students who received the K-PAVE intervention had better vocabulary knowledge than the control group in vocabulary development and academic knowledge (using age-equivalent scores), however there were no differences in listening comprehension. The current study focuses on the student performance measures only: EVT-2, KTEA-II Listening Comprehension subtest, and *D*, measures for expressive vocabulary, listening comprehension, and lexical diversity, respectively. *D* is the primary measure of interest in the current study, and the EVT will be used to examine the relationship and convergent validity between lexical diversity and expressive vocabulary. Similarly, the current study will use the KTEA subtest to examine *D*'s relationship with school skills like listening comprehension.

Statement of Purpose

The primary purpose of the current study is to validate *D* as a useful measure for lexical diversity in at-risk, low-income, predominantly African American children. *D*, if validated on this population, could be a useful supplement to standardized measures of vocabulary which have been shown to be racially and economically biased. Additionally, a valid measure of lexical diversity could provide a more complete picture of a child's language ability by determining the active vocabulary a child has available to use during everyday speech. The use of standardized and nonstandardized measures together could more effectively screen and detect early language deficits in order to determine language ability and guide intervention efforts.

I will use Kane's argument-based approach (1992) to validate *D* as a useful measure of lexical diversity in low-income, predominantly African American young children. Kane's approach "adopts the interpretative argument as the framework for collecting and presenting validity evidence and seeks to provide convincing evidence for its inferences and assumptions, especially its most questionable assumptions" (p. 527). Kane stresses the importance of making inferences from test results based on evidence outlined in interpretative arguments. There are two major components to this approach: the interpretive argument and the validity argument (Kane, 2006). The interpretive argument focuses on the test score as a premise and statements and decisions made from that score as conclusions. The validity argument evaluates the inferences and assumptions made in the interpretive argument using appropriate evidence (Cronbach, 1988). Below is the set of arguments around which this validation is based:

A_1: *D* should yield a range that closely approximates a normal distribution.

A_2: *D* should not be a function of length.

A_3: *D* should have a linear relationship with age.

A_4: *D* should be correlated with other measures of vocabulary, particularly those assessing expressive vocabulary and other production measures.

A_5: *D* should be distinct from TTR and related measures.

The above assumptions used to validate *D* as a measure of lexical diversity are grounded in research regarding the validation of standardized vocabulary assessments in multicultural populations.

Specifically, the current study seeks to investigate the following research questions:

1. Using Kane's argument-based approach, is *D* a valid measure for evaluating lexical diversity in low-income children who are predominantly African American?

2. Is *D* related to measures of school-like tests/skills such as listening comprehension?
3. Is there a difference in *D* between African American children and non-African American children?

CHAPTER 2

METHODS

The current study uses assessment data collected from the K-PAVE study, and therefore a formal request was made to obtain a restricted-use license to receive the dataset. First, a formal request was made through the Institute of Education Sciences (IES) online electronic license application system. After the request was approved by the IES Data Security Office, I completed and returned copies of a signed License (Appendix A), notarized Affidavits (Appendix B), and the Security Plan Form (Appendix C) detailing the procedures to keep the data secure. After approval, the data arrived on CD-ROM encrypted and required a passphrase to open. In order to decrypt the data, an email was sent to the IES Data Security Office for the needed passphrases to open the datasets. Copies of the IES forms are located in Appendices A-C. Along with the Principal Project Officer, Paula Schwanenflugel, I was listed among the personnel allowed to use this data for research purposes.

Sample

The students in this sample are from the Mississippi Delta, a distinctive northwest region of Mississippi. The Delta region was selected for the K-PAVE intervention due to students' increased risk for poor reading outcomes based on the high levels of poverty in the region. Students in the Delta have a history of low achievement scores and the state legislature established a high priority on meeting the early education needs of students in the Delta (Goodson et al., 2010). According to the U.S. Bureau of the Census (2008), the poverty rate in

2007 for children under age 18 in the Delta region averaged 45.3%. The overall poverty rate in 2007 for children under age 18 in the state of Mississippi was 29.7% (18.0% nationally).

The overall analytic sample for the original study contained 1,296 students with 596 students in the intervention group and 700 students in the control group. Differences for dichotomous variables in the original K-PAVE sample were tested using *t*-tests, and results indicated that there were no significant differences between the intervention and control groups for any demographic category. Language samples were collected for approximately half of the total sample from the original study (n = 527). Missing data for *D* scores and EVT standard scores were filled in using stochastic regression by the researchers of the original study. This method is a variant of the regression-based approach in which a stochastic, or random, value is added to the imputed predicted value. These stochastic values are centered at zero, so they do not systematically change the mean; therefore, they provide the same unbiased means as regression imputation (Schlomer, Bauman, & Card, 2010). Adding to regression imputation, stochastic regression corrects for the lack of an error term by adding the average regression variance to the regression imputations to introduce error. Table 2.1 provides the characteristics of the sample that provided language samples.

Table 2.1

Characteristics of Students that Provided Language Samples as Percentages

Characteristics	Intervention Group (*n* = 245)	Control Group (*n* = 282)	Overall Sample (*n* = 527)
Gender			
Female	51.4	51.8	51.6
Male	48.2	47.9	48.0
Race			
African American	85.3	79.8	82.4
Other	14.3	19.9	17.3
Eligibility for free or reduced-price meals			
Yes	93.5	92.2	92.8
No	6.1	7.4	6.8
Has an Individualized Education Program			
Yes	6.6	7.8	7.2
No	93.4	91.8	92.2
Age at post-test			
Mean	73.96 months	73.76 months	73.85 months
Standard deviation	4.530	4.861	4.707

Note. Percentages may not add up to 100% due to missing demographic data. Missing demographic data was removed for final analytic sample.

Data Preparation. After receiving the dataset from IES, a number of steps were taken to reach the final sample for analyses. For this study, students with missing pre- and post-test language samples were removed first. Although *D* scores were imputed using stochastic regression, other language variables (type, token, TTR) were not, and therefore were removed because some analyses are contingent upon those values. Second, any post-test language samples with less than 50 tokens were removed. This was done because the program used to calculate *D*,

vocd, requires a minimum of 50 tokens to obtain *D* values reliably (Malvern et al., 2004). A duplicate ID was also discovered during the data cleaning process and was subsequently removed. Finally, a rubric was developed to remove any student whose transcript contained a high number of issues compared to the whole sample. Each step for determining the final sample and the number of cases removed at each step are listed below in Table 2.2.

Rubric. A rubric was developed to check each transcript to ensure that the validation study would only use high quality child language samples. During the original study, it was reported that during the fall data collection some examiners did not fully understand the protocol. In some instances, examiners were speaking too much during the child's narration, possibly introducing new vocabulary words into the child's language sample. The amount of this kind of talk seemed to vary across transcribers and settings. For this reason, language samples collected in the fall (pre-test) were not used for this validation study. Training for the collection of child language samples was greatly improved and more systematic for the spring (post-test) data collection. Only language samples collected in the spring were used for analyses. I evaluated each student's spring transcript for the quality of the language sample.

The transcribers were trained to code the child's use of new vocabulary words per the CHAT specifications, if the word was introduced by the examiner within one turn. Transcribers were asked to indicate these introduced vocabulary words using the code [word] [/]. Unintelligible speech was coded as 'xxx'. Both were coded to CHAT specifications and, as such, were not included for analyses. For each transcript, the total number of 'xxx' and [word] [/] instances were summed and then divided by the total number of tokens in the sample to obtain a percentage of issues to total number of words. Any transcript that contained 2 *SD*s above the

mean of percent of issues was removed from the sample for analyses. A total of 20 cases were removed from the sample based on the criteria (*M* percent issues = .002; *SD* = .005).

Table 2.2

Steps for Determining Final Sample

Steps	Category	Number of Students Removed	Number of Students Remaining in Sample
	Original Sample		527
1	Students with missing pre- and post-test *D* scores	12	515
2	Language samples with less than 50 tokens	1	514
3	Duplicate IDs	1	513
4	Poor quality language samples, as determined by the rubric	20	493
5	Students with missing post-test data	25	468

Analytic Sample. Tables 2.3 and 2.4 provide information on the final analytic sample after removing each of the cases listed above. Table 2.3 provides a comparison between the treatment groups from the original K-PAVE study and for the final analytic sample in the current study for each level of data (student, classroom, and school levels). Table 2.4 provides the characteristic of the students in the final analytic sample for the current study.

Table 2.3

Comparison of Initial Sample and Final Analytic Sample

Level and Condition	Sample for Original Study	Final Analytic Sample
Schools		
Intervention	30	30
Control	34	34
Classrooms		
Intervention	60	60
Control	68	67
Students		
Intervention	245	224
Control	282	244

Table 2.4

Characteristics of Students in the Analytic Sample

Characteristics	Control (n = 244)	Intervention (n = 224)	Overall (n = 468)
Gender			
Female	132	118	250
Male	112	106	218
Race			
African American	196	191	387
Other	48	33	81
Eligibility for free or reduced-price meals			
Yes	227	209	436
No	17	15	32
Has an Individualized Education Program			
Yes	20	15	35
No	224	209	433
Age at post-test (in months)			
Mean	73.57	73.98	73.76
Standard deviation	4.497	4.486	4.491
Minimum	63	64	63
Maximum	89	90	90

Each analysis conducted to meet the assumptions of the first research question was conducted using only the post-test data from the control group (n = 244). This was done because students who participated in the K-PAVE intervention had known different experiences with vocabulary practices in their classrooms compared to the control group. The intervention had a significant positive effect on vocabulary development and students who received the intervention

had better vocabulary skills than students in the control group as measured by the EVT (Goodson et al., 2010). Further, they had more experience carrying out expressive language as one of the elements of the program, and very close to the outcome measure of interest here (*D* obtained from a language sample).

Measures

Expressive Vocabulary Test – Second Edition. The EVT-2 (Williams, 2007) was the primary measure of vocabulary acquisition in this study. The EVT was standardized on a representative sample of 2,725 examinees ages 2.5 to 90 years, across four U.S. regions. The norming sample included 49.4% females and 50.6% males. Socioeconomic status was assessed based on the education level of the examinee's parents or the education of the examinee, depending on their age. Seventeen percent of the population had less than 12 years of education, 31% were high school graduates, 31% had 1 to 3 years of college or technical school, and 20% had 4 or more years of college. EVT sample distribution by race or ethnicity was 18.1% African American, 12.9% Hispanic, 64.3% White, and 4.6% other. These demographic variables had distributions closely matching those of the U.S. population. Additionally, students receiving various special education services were represented in the norm sample in approximately the same proportion that occurs in the U.S. school population, which included 2.3% students with speech impairments and 5.5% with learning disabilities.

According to the manual, the EVT-2 has internal consistency (split-half) reliabilities of .94-.95 and has a test-retest reliability of .95 (Williams, 2007). The correlations between the EVT-2 and the Peabody Picture Vocabulary Test – Fourth Edition (PPVT-4; Dunn & Dunn, 2007) for children ages 5-6 is .84; correlations for the receptive language, expressive language, and core language scales on the Clinical Evaluation for Language Fundamentals – Fourth Edition

(CELF-4; Semel, Wiig, & Secord, 2003) range between .68 and .80. For more information regarding the specific tasks of the EVT, see the Expressive Vocabulary Test section of the Literature Review chapter of this paper. No independent calculations of validity or reliability for this sample of children were presented in the IES report.

D. *D* is a mathematical calculation of lexical diversity computed by *vocd*, a program that runs within the Computerized Language Analysis (CLAN) program (MacWhinney, 2000) on transcripts that are specifically transcribed in the format of the Child Language Data Exchange System (CHILDES). *D* is calculated by drawing an empirical curve that is compared to a model curve of TTR. Random sampling within the transcript is used in order to best match a single-parameter model and to avoid artificially deflating the score when the speaker remains on the same topic. The procedure for calculating *D* involves beginning at a token size, N, randomly selecting N tokens without replacement from within the sample, and calculating TTR for those N tokens. The tokens are then replaced within the sample, and this process is repeated 100 times. The average TTR is calculated for that N, and then the next number of tokens (N + 1) is chosen and the entire procedure begins again. Once these calculations have been accomplished for the entire range of tokens, 35–50, each point (16 points) is plotted and a value *D* is calculated. *D* is the parameter that is adjusted to estimate the best fit of the empirical curve (obtained using real data) to the model curve through the least squares method. The value *D* is calculated three times through this procedure, and the average is then reported as the final optimum *D* value for that transcript. Since repeated random sampling is used, the entire transcript is sampled, despite the fact that the curve is only estimated from 35 to 50 tokens. Additionally, it is important to realize that a slightly different value of *D* will be reported each time *vocd* is ran, since the value is based on averages of random sampling (McKee et al., 2000; Richards & Malvern, 1997). Although the

program *vocd* will output *D* values even for extremely small language samples, Malvern et al. (2004) recommended that samples contain at least 50 tokens to produce reliable values of *D*.

Kauffman Test of Educational Achievement – Second Edition. The Listening Comprehension subtest of the KTEA-II (Kauffman & Kauffman, 2004) was used to assess listening ability and understanding. The Listening Comprehension is a subtest for the Oral Language factor of the KTEA. This subtest measures a student's ability to listen, without assessing reasoning skills or memory of details. The examinees listen to a passage on an audio CD and then are asked questions about the story by the examiner. The subtest contains 18 passages and 61 questions, which include multiple-choice and open-ended factual questions. The examiner may repeat questions, but each passage may only be played once. The KTEA-II was standardized with an age-norm sample of 3,000 examinees ages 4½ to 25 and a grade-norm sample of 2,400 examinees in kindergarten through grade 12. The sample was matched to the March 2001 U.S. Census Bureau statistics for gender, geographic region, education level of examinee's mother, ethnicity, and parental education within each ethnic group. Although the demographics for the norming sample were not included in the KTEA manual, the March 2001 U.S. Census Bureau Annual Demographic File (U.S. Census Bureau, 2001) reports that 65.4% of the population was African American (if using the unit count) or 71.6% of the population was African American (if using the weighted count).

According to the manual, the Oral Language composite reliability tended to be lower than other areas of the test (.87). The Listening Comprehension subtest of the KTEA-II has an average internal consistency (split-half) reliability of .85 for both grade-level and age-level. For kindergarteners, the Listening Comprehension subtest has a split-half reliability of .84. Results from a confirmatory factor analysis with students in grade 1 and higher indicate that the

correlation between the Oral Language factor (listening comprehension and oral language subtests) and the Reading factor is .91 and the errors for the Listening Comprehension subtest and the Reading Comprehension subtest are correlated. The correlation in grades 1-5 between the KTEA-II Listening Comprehension subtest and the Woodcock-Johnson III Listening Comprehension test is .77.

Materials

Prompting pictures. During the original study a number of pictures were presented to the child prior to narrating the picture book. These pictures served as a warm-up and were used so the child could familiarize themselves with the researcher and become comfortable narrating the story. A total of 7 pictures were available, although not all of them were presented and discussed with each child. Pictures included images of McDonalds, Santa Claus, a child at the dentist, a bee on a flower, a doctor checking a child's heart, a child wearing a cast on her leg, and children exiting a school bus.

Picture books. Children were assigned to narrate one of two different wordless picture books in a series by Alexandra Day: *Carl Goes Shopping* (Day, 1992) and *Good Dog Carl* (Day, 1996). *Carl Goes Shopping* tells the story of a Rottweiler, Carl, who takes care of a baby in a shopping mall while the mother goes shopping. *Good Dog Carl* tells the story of Carl, who takes care of a baby at home while the mother is away. The books are illustrated in color and only contain text on the first page to set up the story. The books contain between 27 and 31 pictures and have a clear story line with an obvious sequence of events. These wordless picture books have been used in previous research studies investigating basic language skills in preschoolers (e.g., Ruston & Schwanenflugel, 2010; Rvachew, Ohberg, Grawburg, & Heyding, 2003).

Analytical Software

Computerized Language Analysis. Computerized Language Analysis (CLAN) is a program that was designed to analyze data transcribed in the format of Child Language Data Exchange System (CHILDES). CHILDES developed the computational tool that, among other things, automates the process of data analysis in the form of transcript data. Transcripts were transcribed in CHAT mode, the simpler of two modes used for basic editing commands. CLAN allows you to perform a large number of automatic analyses on transcript data including frequency counts, word searches, co-occurrence analyses, mean length of utterance (MLU) counts, interactional analyses, text changes, and morphosyntactic analysis (MacWhinney, 2000). In this case, CLAN was used to run the program, *vocd*, in order to calculate four language measures: type, token, type-token ratio, and *D*. Each value was then recorded in a log. To see a sample transcript and CHAT specifications see Appendix D.

IBM SPSS Statistics. SPSS version 21 was used to analyze the dataset provided by the Institute of Education Sciences (IES) for the K-PAVE project. The IES dataset was provided in SAS format, but was converted to SPSS for analysis. All statistical analyses carried out in the current study were performed using SPSS.

Procedures

Language samples were collected using a digital audio tape recorder. The examiner escorted the child to a quiet area in a hall or room immediately outside of the classroom. The examiner engaged the child in conversation by prompting them with a series of photos intended to elicit personal narratives and make the child feel comfortable with the examiner and task. Children were then asked to narrate a wordless picture book, *Good Dog Carl* or *Carl Goes Shopping*. This was counterbalanced so that if the child received one book for pre-test, he or she

received the other for post-test. When introducing the book, the examiners explained that there was no right or wrong story, but the child might make up any story to go along with the pictures. As per a protocol developed by Ruston and Schwanenflugel (2010), examiners were instructed to use only the following prompts to limit the amount of speech that the examiner introduced into the sample: (1) Can you tell me more?, (2) What's happening?, (3) What else do you see? Any language samples with additional prompts were reviewed more carefully. The goal behind these prompts was to help the examiners extend the conversation about the book while avoiding the introduction of new vocabulary terms.

Language samples were transcribed using the program Transcriber 1.5.1, a tool used for segmenting, labeling, and transcribing speech (Barras, Geoffrois, Wu, & Liberman, 2001). Language samples were transcribed according to CHAT specifications. See Appendix D for an example transcript written in CHAT specifications and Appendices E and F for CLAN commands and output (for FREQ and VOCD commands, respectively). Transcriptions were carried out by one of 10-15 trained transcribers and were rechecked to ensure transcription accuracy.

Transcriber training. All transcribers had either extensive knowledge of African-American Vernacular English (AAVE) or child language acquisition through experience (such as living in communities where a vast majority were African American, or having gone to schools having heavy representation of African Americans, or being of African American descent) or having substantial coursework in which dialect issues were discussed (such as linguistics or speech and language disorders). To ensure a common understanding among transcribers, all transcribers were trained to recognize common features of child language for five-year-olds, particularly those found in African-American dialect. All child language samples were

transcribed by a transcriber trained to carry out the analysis. The first 10-15 samples were listened to and re-checked by the trainer, a graduate student, and fixed by the trainee until the trainee transcriber had full command of the procedures.

Transcripts. Child language samples were transcribed using the program Transcriber 1.5.1. Words, as indicated by Merriam-Webster, were transcribed in lower-case while proper nouns were transcribed appropriately. Any attempt on the part of the interviewer to have the child continue the protocol beyond the narration of the wordless storybook was eliminated. When a word or phrase was unintelligible, the transcribers indicated this by transcribing 'xxx', the code specified by CLAN, which was excluded during analysis. A second, more experienced transcriber with in-depth familiarity with AAVE in kindergarten children reviewed the segments indicated by these missing elements and filled in the unclear speech if it could be understood. If not, the 'xxx' code was left in and excluded from the analysis. Each transcript was also checked for spelling and grammatical errors before running analyses.

Quality control. A random 1½ minutes of each recording was listened to by a more experienced transcriber to determine if the first transcriber had adequately transcribed the child narrative. One and a half minutes represents 10-15% of the median recording time (10:56) for the spring samples. During the 1½ minutes any errors that were made were corrected, including misunderstandings of vocabulary, misspellings of words, and general grammatical errors (i.e., adding periods or eliminating apostrophes for analysis). If the difference (or changes that needed to be made) were greater than 8 words then the transcript was assumed to be inadequate and the entire transcript was transcribed again to correct errors throughout. The criterion of 8 words represents 1% of the average child tokens across all the recordings.

The transcripts were then scanned for words introduced by the interviewer. The word is considered a newly introduced word if the interviewer mentions a vocabulary word that was not previously said by the child (e.g., Have you ever had a *cast* on before?). Clark (2007) indicates that children who are going to uptake new words from adult speech generally do so within the next turn. Consequently, if the child said a content word (nouns, verbs, adjectives, and adverbs) that was introduced by the interviewer first, this word coded as [word] [/] in the transcript and excluded from analyses. Using this helped control for variations among testers in the introduction of potential new words from children's speech. A second spell check was done to ensure that all words were spelled correctly and all grammatical errors were eliminated before re-analyzing.

Reliability. Nine percent of each transcriber's total completed transcripts were re-transcribed by another transcriber to measure reliability. A total of 44 transcripts collected in the spring were transcribed for reliability. Analyses on all four measures (type, token, type-token ratio, and *D*) were compared and transcripts were considered reliable if it produced a correlation of greater than .80. A correlation addressed the reliability between a transcript produced by transcriber 1 and transcriber 2. Type, token, type-token ratio, and *D* showed to be highly correlated for the spring samples between transcriber 1 and transcriber 2, $r(42) = .99$, $p = .05$, $r(42) = .98$, $p = .05$, $r(42) = .97$, $p = .05$, and $r(42) = .98$, $p = .05$, respectively.

CHAPTER 3

RESULTS

The results are analyzed and presented in order of research question. The first several analyses were conducted to address the first research question of whether *D* is a valid measure for evaluating lexical diversity in low-income, predominantly African American children. In order for *D* to be a valid measure for evaluating lexical diversity in this population, several assumptions were proposed and tested, therefore, each section of the results section will address the assumptions for each research question. Each analysis conducted to meet the assumptions of the first research question was conducted using only the post-test data from the control group (n = 244). This was done because the intervention and control groups had known different experiences with vocabulary practices in their classrooms. From the original K-PAVE study, it was reported that the intervention had a significant positive effect on vocabulary development and that students who received the intervention were one month ahead of students in the control group as measured by the EVT (Goodson et al., 2010).

Normality

The first assumption proposes that *D* scores should yield a range that closely approximates a normal distribution. Various tests for evaluating normality in *D* scores were conducted. Review of the Shapiro-Wilks test for normality (SW = .978, df = 244, p = .001) suggests that the distribution of *D* values are non-normal. However, the Shapiro-Wilk's test is stringent and may be too sensitive in detecting non-normality for large data sets, thus skewness and kurtosis values were also evaluated. Kim (2013) suggests obtaining a z-score by dividing the

skewness and kurtosis statistics with their respective standard errors to evaluate normality for samples of greater than 50 but less than 300. A quotient (z-score) of less than 3.29 suggests the data are normal. Using Kim's suggestion, this yields a value of 3.72 for skewness (skewness = .581, *SE* = .156) and 2.11 for kurtosis (kurtosis = .656, *SE* = .310), and suggests the data are skewed. A positive skewness statistic indicates that the data are skewed to the right (right side of the distribution is longer than the left side), with most values concentrated to the left of the mean and extreme values to the right. Figure 3.1 presents a histogram with the normal distribution curve overlay. Figure 3.2 presents a box-and-whisker plot with outliers. Although there are some outliers, they were left in the sample to represent true variation in the data.

Figure 3.1

Histogram with Normal Distribution Curve Overlay

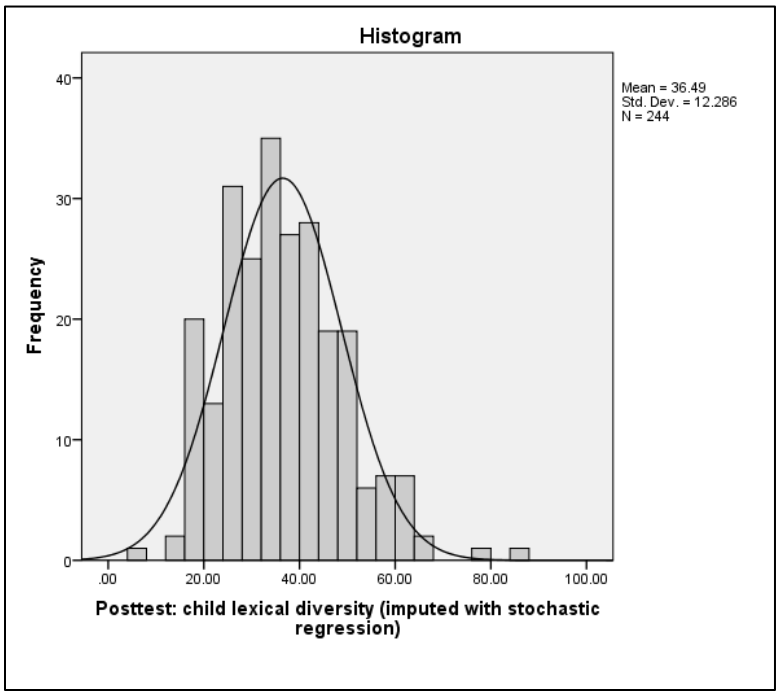
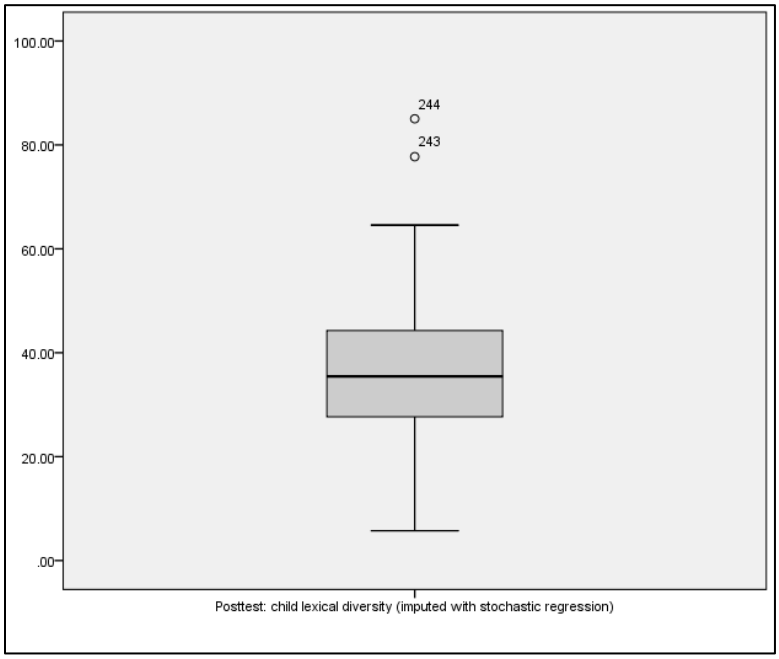


Figure 3.2

Box-and-Whisker Plot with Outliers



Sensitivity to Sample Size

The second assumption proposes that *D* scores should not be a function of length. Malvern and colleagues (2004) claim that *D* is not a function of length. To investigate *D*'s sensitivity to sample size I used the split-half facility in the *vocd* program of CLAN. The split-half facility provides *D* values for even-numbered and odd-numbered words, allowing a comparison between each to the *D* of the whole transcript. In other words, the *D* values for half the words are compared to the *D* value for all of the words in each transcript. Type-token ratio was also examined for even-numbered and odd-numbered words. It has been shown that higher values of TTR are obtained from shorter samples and similarly, lower TTR values from larger samples (Malvern et al., 2004). Further, if the topic of speech is focused, then it is likely that the speaker will reuse content words related to that topic thus decreasing TTR (by increasing number of tokens but not type). Calculating split-half reliability by analyzing even- and odd-numbered

words minimizes local contextual effects or the number of topic changes, focusing on whether the scores are a function of the size of the sample.

Means for even-numbered words, odd-numbered words, and whole transcript are reported in Tables 3.1 and 3.2 below for *D* and type-token ratio, respectively. Results from *t*-tests comparing even-numbered and odd-numbered words to the whole transcript are also reported in the tables below.

Table 3.1

D for Different Parts of the Transcript Compared to Whole Transcript (n = 243)

	M	*SD*	*df*	*t*	*p*	*d*
Even-numbered words	36.67	12.80	242	.172	.864	.011
Odd-numbered words	36.86	13.76	242	.771	.442	.049
Whole transcript	36.61	12.15				

Note. One transcript was unable to be analyzed using the split-half facility because once split in half the transcript did not have enough tokens for the random sampling process used by the program *vocd*.

Table 3.2

TTR for Different Parts of the Transcript Compared to Whole Transcript (n = 243)

	M	*SD*	*df*	*t*	*p*	*d*
Even-numbered words	0.413	0.069	242	70.25	< .01	4.506
Odd-numbered words	0.412	0.075	242	67.39	< .01	4.323
Whole transcript	0.314	0.064				

Note. One transcript was unable to be analyzed using the split-half facility because once split in half the transcript did not have enough tokens for the random sampling process used by the program *vocd*.

As hypothesized, and consistent with the developers' claim, *D* is not a function of sample size. There are no significant differences between *D* of even-numbered words when compared to *D* of the whole transcript. Likewise there are no significant differences between *D* of odd-numbered words when compared to *D* of the whole transcript. These findings also support previous findings by Silverman and Ratner (2002) and Wright, Silverman, and Newhoff (2003) who used a similar method by examining the sensitivity of *D* for even-numbered and odd-numbered utterances in children who stutter and aphasic patients, respectively. By contrast,

Owen and Leonard (2002) examined *D*'s sensitivity to sample size by comparing the first half of a transcript to the whole transcript. Using this method, Owen and Leonard found that *D* values were significantly different for the two sample sizes. This method of comparing the first half of a transcript to the whole does not control for the greater number of topic changes likely to occur in the larger sample. Analyzing every other word throughout the transcript is preferable to dividing transcripts into two halves or sections. Additionally, the present findings are consistent with previous findings (Malvern et al., 2004), where unlike *D*, TTR is significantly affected by sample size. Type-token ratio for half of the transcript is significantly higher than the TTR for the whole transcript, confirming that shorter transcripts produce higher ratios than longer transcripts where words are likely be repeated.

Internal consistency of transcripts was also evaluated using odd- and even-numbered words compared to the whole transcript. Table 3.3 presents correlations of *D* between the whole transcript and the two halves.

Table 3.3

Correlations between D for Whole Transcript and Even- and Odd-numbered Words (n = 243)

	Even-numbered Words	Odd-numbered Words
Whole transcript	.925*	.934*
Odd-numbered words	.756*	

Note. *$p < .01$

The results indicate that *D* for both odd- and even-numbered words significantly correlates to *D* for the whole transcript. Because the number of tokens was split in half, the correlation of *D* between the two halves of the transcript is lower than the correlations between each half and the whole transcript. One problem with the split-half reliability coefficient is that it is based on half of the full transcript and the reliability coefficient is generally reduced with smaller sample sets. Therefore, the Spearman-Brown correction (Crocker & Algina, 1986) was

applied to adjust for the shortened transcripts and the internal consistency estimate between the odd- and even-numbered words was r_{sb} = .861, indicating satisfactory reliability.

D's Relationship to Age

The third assumption proposes that *D* should have a linear relationship with age. To evaluate if *D* has a relationship with age, *D* and age were correlated. The results indicate that *D* does not correlate with age, $r(242) = .069$, $p = .28$. This is in contrast to the developer's claims (Durán et al., 2004) and previous findings (Klee et al., 2004) which suggest that *D* has a linear relationship with age. This contrasting finding could be attributed to the differences in ages and age ranges between the studies. Whereas the ranges in studies by Durán and colleagues and Klee and colleagues ranged from 27-68 months and 18-60 months (with at least 10 students per age group), respectively, the ages in the current study only ranged from 65-81 months for months containing more than one case. A relationship between age and *D* may not be evident in older children unless there is a wider range of ages.

D's Relationship to Other Measures of Vocabulary

The fourth and fifth assumptions are presented in this section together since they are both related to *D*'s relationship with other measures of vocabulary. The fourth assumption proposes that *D* scores should be correlated with other measures of vocabulary, particularly those assessing expressive vocabulary. Table 3.4 presents the descriptive statistics for all language variables.

Table 3.4

Descriptive Statistics of Language Variables (n = 244)

Variable	*M* (*SD*)	Min	Max	Skewness (*SE*)	Kurtosis (*SE*)
Type	164.94 (58.64)	26	435	0.828 (0.156)	1.648 (0.310)
Token	566.18 (282.74)	56	2072	1.338 (0.156)	3.455 (0.310)
Type-Token Ratio	0.315 (0.064)	0.186	0.674	1.381 (0.156)	4.215 (0.310)
D	36.49 (12.29)	5.74	85.03	0.581 (0.156)	0.656 (0.310)
EVT	92.13 (11.31)	62	127	0.214 (0.156)	0.069 (0.310)

To evaluate if *D* has a relationship with other measures of vocabulary, particularly those assessing expressive vocabulary, *D* scores were first correlated with EVT standard scores. It was hypothesized that there should be a moderately strong correlation between these measures of expressive vocabulary. *D* was significantly correlated with EVT standard scores, $r(242) = .24$, $p < .01$. The correlation between *D* and EVT standard scores is weaker than previous findings which found *D* to have a moderately strong correlation ($r = .48$) to Expressive One-Word Picture Vocabulary Test – Revised, another standardized measure of expressive vocabulary (Silverman & Ratner, 2002).

D scores were also correlated with other language measures produced by the *vocd* program in CLAN. *D* was correlated with type (number of unique words). There was a strong correlation between *D* and type, $r(242) = .665$, $p < .01$. This is consistent to the previous findings by Malvern and Richards (2002). Correlations between *D* scores and tokens (total number of words) and *D* scores and TTR were also calculated. It was hypothesized *D* would be more strongly correlated with EVT and type, consistent with previous findings, than either tokens or

TTR. In fact, *D* was moderately correlated with tokens, $r(242) = .514$, $p < .01$, but did not correlate with TTR, $r(242) = 0.016$, $p = .801$. Although *D* and tokens were correlated, the correlation between *D* and type was greater. Because *D* measures how diversely vocabulary resources are deployed, it is not surprising that *D* would have a strong correlation with more similar measures of productive language, such as type, than to standardized assessments.

The fifth assumption proposes that *D* should be distinct from TTR and related measures. That is, correlations between EVT and type with *D* should be higher than correlations between token and TTR with *D*. To test if there are significant differences between correlations of *D*, *z*-scores were calculated for each pair of significant correlations. Specifically, the test was done for the following pairs of correlations: (1) *D* and token and *D* and EVT; (2) *D* and EVT and *D* and type, and (3) *D* and type and *D* and token. This procedure was not done for *D* and TTR since these variables are not significantly correlated. FZT Computator was used to calculate *z*-score. First, the correlations between xy, xz, and yz are entered into the computator (for example, *D* and EVT, *D* and token, and EVT and token). The sample size is also entered into the computator to calculate degrees of freedom. Then, each correlation coefficient is converted into a *z*-score using Fisher z-transformation applied to the sample correlation coefficient *r*. A *z*-score is calculated and by using the degrees of freedom, the critical values of *t* can be looked up. The correlations of token and EVT with *D* are significantly different, $Z = 4.037$, $p < .01$, indicating that the correlation between token and *D* is significantly greater than the correlation between *D* and EVT. The correlations of type and EVT with *D* are also significantly different, $Z = 6.979$, $p < .01$, indicating that the correlation between type and *D* is also significantly greater than the correlation between *D* and EVT. These results indicate that although *D* is significantly correlated to EVT, token, and type, the correlations between *D* and token and *D* and type (similar measures

of productive language analyzed in CLAN) were also significantly different than the correlation between *D* and EVT. Further, the correlations of type and token with *D* were significantly different, $Z = 9.675$, $p < .01$, indicating a significantly stronger correlation between *D* and type than *D* and token, as hypothesized. These results indicate that *D* is most strongly related to type than the other language variables. *D*'s highest correlation, with type, indicates that these two measures are closely measuring the same construct. *D* and EVT, although significant, are weakly correlated, suggesting that *D* is measuring something unique to the EVT.

The first research question sought to determine if *D* a valid measure for evaluating lexical diversity in low-income, predominantly African American, children. This was determined by testing assumptions using Kane's argument-based approach to validity. The results were mixed, with some assumptions being met while others were not. The first assumption, that the data should be approximately normally distributed, was met, although the data were slightly skewed due to a few extreme values. The data did yield a range of scores (min = 5.74, max = 85.03) and generally approximated a normal distribution. The second assumption that *D* is not a function of sample size was also met, as there were no significant differences between *D* scores for half of the transcript (by analyzing even- and odd-numbered words) compared to the whole. The third assumption was not met, as *D* did not have a linear relationship with age. This could be due to the narrow range of ages in this study. In analyzing for the fourth and fifth assumption, I found that *D* has the strongest relationship with type, or number of unique words in a language sample. Although *D* did have a significant correlation with EVT, *D* was more strongly correlated with type and token. This indicates that *D* is most strongly correlated with other similar production measures, more so than a standardized assessment of expressive vocabulary, suggesting it is measuring something unique.

D's Relationship to Listening Comprehension

The second research question addresses *D*'s relationship with school-like skills such as listening comprehension. For young children, listening comprehension is an understanding of stories and other texts that are read aloud to them. Listening comprehension provides the foundation for children to be able to understand what they've read, remember what they've read, and communicate that with others. To evaluate if *D* has a relationship with listening comprehension, *D* and the standard scores on the Listening Comprehension subtest of the KTEA-II were correlated. Similar to the sample used to address the first research question, only the post-test data from the control group was analyzed because of the different experiences with vocabulary practices between the control and intervention groups. Overall, the average KTEA score for the Listening Comprehension subtest was 88.92 (*SD* = 12.86). The correlation indicates that *D* is significantly, but weakly, correlated with listening comprehension, $r(242) = .16$, $p = .013$. I also conducted a one-way ANOVA to see if *D* predicted listening comprehension, as measured by the KTEA subtest. Results indicate that *D* does not significantly predict KTEA Listening Comprehension scores, $F(1, 235) = 1.509$, $p = .276$, $\eta_p^2 = 978$. Although there is a small yet significant correlation between *D* and listening comprehension, *D* does not predict it.

Multiple regression analysis was used to test if *D* accounts for any additional variance in listening comprehension once EVT scores are taken into account; that is, this analysis sought to determine if *D* adds to our ability to predict listening comprehension. The data were screened for violation of assumptions prior to analysis and assumptions of linearity, normality, independence, homogeneity of variance, and multicollinearity were met. The results of the regression indicated that the two predictors, EVT and *D*, significantly explained 48.8% of the variance ($R^2 = .488$, $F(2, 241) = 114.69$, $p < .01$). When evaluating each individual predictor, it was found that EVT

significantly predicted KTEA scores (β = .700, $p < .01$), accounting for 46.2% of the variance, however *D* did not significantly contribute to the model (β = -.008, $p = .861$).

Differences between African American and Non-African American Children

A secondary focus of this study was to examine if there are any performance differences between African American children and non-African American children. Previous research investigating performance differences between these populations have found that African American children perform lower on standardized assessments of vocabulary compared to European Americans (Qi et al., 2006, Restrepo et al., 2006; Webb et al., 2008). To investigate if there are any differences between African American children and non-African American children, an independent samples *t*-test was conducted on the following measures: *D*, EVT standard scores, and standard scores from the KTEA Listening Comprehension subtest. Similar to the sample used to address the first and second research questions, only the post-test data from the control group was analyzed because of the known differences with in-classroom vocabulary practices between the control and intervention groups.

Because the sample sizes were uneven between children who were categorized as African American ($n = 196$) and non-African American ($n = 48$), Levene's Test for Equality of Variances was used to assess the assumption that the population variances are equal (i.e., homoscedasticity). The Levene's tests were nonsignificant for all three measures, indicating that equal variances between groups can be assumed. Table 3.5 shows the descriptive statistics for each of the three measures and *t*-test results between the two groups.

Table 3.5

Performance Differences between African American and Non-African American Children

	African American		Non-African American				
Variable	*M*	*SD*	*M*	*SD*	*t*(242)	*p*	*D*
D	36.28	12.22	37.35	12.66	0.541	.589	-0.086
EVT	91.23	11.17	95.79	11.25	2.53	.012*	-0.407
KTEA	88.13	12.81	92.15	12.70	1.95	.052	-0.315

Note. *$p < .05$

Consistent with previous research, the findings indicate that there were significant differences in performance on the EVT between the African American children and non-African American children. Further, a one-sample *t*-test revealed that the mean scores for both groups were significantly lower than the standardized mean of 100, $t(195) = -10.99$, $p < .01$ and $t(47) = -2.593$, $p = .013$, for African American and non-African American children, respectively. These results could indicate that children of low SES perform lower on standardized assessments than the norming sample, with African American students further performing lower. Similarly, there was a marginal difference between African American and non-African American performance on the KTEA. Like the EVT, the mean scores for the KTEA for African American and non-African American children were significantly lower, $t(195) = -12.971$, $p < .01$ and $t(47) = -4.286$, $p < .01$, respectively, than the standardized mean of 100. Further, African American children performed lower than non-African American children. Unlike the EVT and KTEA, there were no significant differences on *D* between the two groups, suggesting that *D* may not suffer from racial bias like standardized assessments have been shown to exhibit. However, as a whole, *D* scores are lower in this sample than samples of similar ages from other studies. Durán and

colleagues (2004) found that the mean *D* score in their sample of 5-year-olds was 64.02, which is much higher than the 36.49 for this sample of 6-year-olds. This suggests that students from low-income backgrounds are less lexically diverse than children from middle SES, but *D* may be a fairer measure, by alleviating racial bias.

CHAPTER 4

DISCUSSION

The primary purpose of the current study was to determine if *D* is a useful measure for lexical diversity for children in schools where the majority of children are at-risk for low academic performance, low-income, and predominantly African American. The current study adds to existing literature by examining the validity of *D* in this specific population. This was examined using Kane's argument-based approach to validity (1992). As noted earlier, there are two major components to Kane's approach: the interpretative argument and the validity argument. The interpretive argument includes the assumptions and inferences leading to the statements and decisions that can be made from assessment results (i.e., the interpretation of test scores). The validity argument evaluates the interpretative argument as a whole and provides the rationale for accepting the inferences and assumptions included in the interpretation. In addition to the validation of *D*, the current study also sought to examine if *D* is related to measures of school-like tests/skills such as listening comprehension. Lastly, the current study examined if there are any differences between African American children and non-African American children from these kinds of schools when measuring lexical diversity with *D*. The findings regarding each of the research questions will be discussed below in sequence.

The Validity of *D* Using Kane's Argument-Based Approach

The findings from this study provide some evidence in support of *D* as a valid measure for evaluating lexical diversity in low-income children who are predominantly African

American. Using Kane's argument-based approach (1992), five arguments were proposed around which this validation is based.

A_1: *D* should yield a range that closely approximates a normal distribution. The first assumption, that the data should approximate a normal distribution, was met. A test that is appropriate for a specific population, such as African American students, should yield a range of performances from well above average to below average, closely approximating a standard normal distribution (Washington & Craig, 1999). Although the data exhibited some skewness, this was only due to a few extreme cases. The *D* scores ranged from 5.74 to 85.03, a wide range spanning approximately -2 *SD* below to mean to +2 *SD* above the mean. This suggests a sufficient performance spread on this measure and thus *D* is appropriate for representing diversity in expressive use of vocabulary among low-income predominantly African American children.

A_2: *D* should not be a function of length. The second assumption, that *D* should not be a function of length, was met. Previous measures used to quantify lexical diversity such as number of different words (NDW) and type-token ratio (TTR) have been criticized because of their dependence on sample size. Longer transcripts will produce higher NDW than shorter transcripts, as different words are introduced as language samples get longer. In contrast, shorter transcripts will produce higher TTR than longer transcripts, as words would be repeated as language samples increase driving the TTR down. *D* was developed specifically to overcome the sample size issue (Malvern et al., 2004). To assess the construct validity of *D*, scores for even- and odd-numbered words were compared to *D* scores for the whole transcript; there were no significant differences found between the halves of the transcript and the whole. With correlations above 0.90, the *D* scores from odd- and even-numbered words compared to the

whole transcript are highly reliable. These findings suggest that *D* does overcome the dependency to sample size.

A_3: *D* should have a linear relationship with age. The third assumption, that *D* should have a linear relationship with age, was not met. Correlations between the two variables indicated that there is no significant relationship between *D* and age among children in this age range. This is in contrast to previous findings (Durán et al., 2004; Klee et al., 2004) which suggest that *D* has a linear relationship with age. This was somewhat surprising given the high correlations between the two in previous studies, but the difference in findings might be attributed to the ages of the samples and the limited SES range of the current sample. Previous studies have had a wider range of ages in their sample than the current study. Whereas the current study assessed children in kindergarten (range: 16 months, or between ages 65-81 months), samples from previous studies had age ranges of more than three years. Further, previous studies have included children as young as 18 months. Between 18 months and school-aged, children's vocabulary growth is steep; typical developing children acquire over a thousand words between those ages (Stahl, 1999). A relationship between age and *D* may not be as evident in older children unless wider ranges of ages are included in the sample. Furthermore, almost 100% of the sample was considered low SES, and a limited range in SES may also contribute to the lack of correlation between age and *D* scores.

A_4: *D* should be correlated with other measures of vocabulary, particularly those assessing expressive vocabulary and other production measures. The fourth assumption, that *D* should be correlated with other measures of vocabulary, was met. To assess the convergent validity of *D*, it was hypothesized that there should be a moderately strong correlation between these measures of expressive vocabulary. *D* was significantly correlated with the EVT; however,

the correlation was weaker than previous findings which found *D* to have a moderately strong correlation with another standardized measure of expressive vocabulary, the EOWPVT-R (*r* = .48, Silverman & Ratner, 2002). Additionally *D* was strongly significantly correlated with type and token, number of unique words and number of total words in a language sample, respectively. Because *D* measures how diversely vocabulary resources are deployed, it is not surprising that *D* would have a stronger correlation with more similar measures of productive language than to standardized assessments.

A_5: *D* should be distinct from TTR and related measures. The fifth assumption, that *D* should be distinct from other measures, was met. *D* was not significantly correlated with TTR, supporting previous findings (Malvern & Richards, 2002; Silverman & Ratner, 2002). Further, in testing differences between significant correlations, I found that *D* was most significantly correlated with type and therefore is most strongly related to number of different words in a language sample than to other language variables such as token or EVT scores. Lexical diversity, as measured by *D*, is measuring something unique to expressive vocabulary such that the deployment of active vocabulary is different from knowing meanings or synonyms of target words.

The above results validate *D* and suggest that *D* could be a useful and informative measure of lexical diversity in schools having predominantly low-income, at-risk, African American children. *D* could be a useful supplement to standardized measures of vocabulary which have been shown to be racially and economically biased. *D* could provide a more complete picture of a child's language ability by determining the active vocabulary a child has available to use during everyday speech. The EVT is a widely-used measure expressive vocabulary; however, the target words and definitions on the test don't necessarily represent a

child's lexical range, especially if those target words are unfamiliar to the child. Words that are acquired as part of one's vocabulary are influenced by cultural experiences. The meanings of words can vary substantially for people from different linguistic communities, and people from the same linguistic community can have multiple meanings for words (Stockman, 2000). *D* provides an alternative measure to expressive vocabulary deployment that is not linked to a child's knowledge of specific vocabulary words. The use of standardized and nonstandardized measures together could more effectively screen and detect early language deficits in order to determine language ability and guide intervention efforts. In fact, previous research has suggested that used alone, the EVT may not be able to identify language disorders in African American children (Thomas-Tate et al., 2006). Similarly, the PPVT was also found to be inadequate for diagnosing language disorders in this population (Campbell et al., 2001), which tended to underestimate achievement in this population. Further, the relatively small, but significant, correlation between *D* and EVT suggests that *D* is measuring a different dimension of vocabulary, possibly a more communicative dimension of expressive vocabulary that is not limited only to definitions and synonymy.

D and Listening Comprehension

Similar to vocabulary knowledge, listening comprehension skills are important to reading comprehension. In fact, listening comprehension skills have been found to strongly relate to and predict reading comprehension skills (Kendeou, van den Broek, White, & Lynch, 2007; Nation & Snowling, 2004). Listening comprehension is especially important in kindergarten because these children are not yet proficient at reading; kindergarteners rely on their listening comprehension skills to acquire new vocabulary and to comprehend texts. Given the strong relationship between reading comprehension and listening comprehension, it was important to

investigate *D*'s relationship with listening comprehension in addition to expressive vocabulary skills. Similar to expressive vocabulary, *D* was significantly, although weakly, correlated with listening comprehension. *D* is more strongly related to expressive vocabulary than listening comprehension, which is not surprising because *D* is the deployment of expressive vocabulary. However, the current study suggests that, if the goal of the assessment is to obtain a measure of vocabulary that will predict reading comprehension, it might be better to use the EVT rather than *D*.

Differences between African American and Non-African American Children

A comparison between *D* in African American and non-African American children was a secondary focus of this study. Previous research in standardized vocabulary tests found that these measures can be racially-biased. One of the criticisms of standardized vocabulary tests is the nature of the tasks; the tasks may be unfamiliar and too decontextualized. African American children perform better when test items are provided in context of thematic activities (e.g., a story or arts and crafts) compared to traditional administration (Fagundes et al., 1998). The current study found that there were no differences in *D* between African American and non-African American children. This suggests that *D* may not suffer from racial bias the way that standardized assessments of vocabulary have been shown to exhibit. The results for *D* are unlike the EVT which did find significant differences in performance between these two groups. Although the mean EVT scores for both groups were lower than the standardized mean, the mean EVT scores for African American children were still lower than the mean scores for non-African American children. The results from the current study adds to previous studies that also found performance differences in these two groups on the EVT (Qi et al., 2006; Restrepo et al., 2006, Webb et al., 2008).

For both EVT and KTEA, children in this sample performed lower than the standardized mean, with the African American children further performing lower than the non-African American children. More than 90 percent of the children in this sample were eligible for free or reduced lunch; these results support previous findings that low-income children perform lower on standardized measures than children from middle SES (Restrepo et al., 2006). Further, although *D* did not seem to exhibit racial bias, as a whole, scores were lower than other studies with similar-aged children (Durán et al., 2004). This is not surprising since the children in the Mississippi Delta were targeted for the vocabulary intervention due to their low student achievement. It seems that *D* may be a fairer since it alleviates the racial bias that has been exhibited in standardized tests.

The process used to collect language samples may have limited the speech produced by the sample. Again, rather than spontaneous speech language samples, children were asked to narrate a wordless picture book, perhaps constraining the vocabulary expressed. However, the non-difference between African American and non-African American children was not result of the picture books selected, *Good Dog Carl* and *Carl Goes Shopping*. In fact, Ruston (2007) used the same wordless picture books and found that in four-year-olds the average *D* score was 41.15 (*SD* = 15.68), suggesting that *D* scores in the current study were not constrained by using these particular books. However, Ruston did find that when compared to language samples derived from personal narratives (various prompts were used to elicit spontaneous speech about children's personal lives) *D* scores were significantly higher for personal narratives (*M* = 55.46, *SD* = 10.39). This suggests that *D* scores may be higher for spontaneous speech than speech directed by narrating specific books and/or story lines.

Limitations of *D*

There are several limitations in obtaining *D* scores. First, calculating *D* is a time consuming and laborious task. The user must obtain a language sample with enough substance to calculate *D* and transcribe it in accordance with the specifications of the *vocd* program. The amount of time it takes to obtain *D* may not be practical for all practitioners (e.g., school teachers, speech pathologists, school psychologists) considering the brevity of some standardized measures of vocabulary. Some of this time and labor can be alleviated by using automated transcription software designed for natural language speech recognition such as MacSpeech Scribe or PowerScribe. Second, there is no standardized procedure for collecting language samples (or writing samples). For pre-readers, the protocol employed in this current study seems appropriate. Having children narrate a wordless picture book at least standardizes the amount of pages a child sees to narrate; the number of pages in the book is held constant for each child. Although the amount of time narrating and the number of words spoken vary by child, fixing the number of pages and prompts by the examiner is one suggestion for standardizing the protocol for obtaining *D*. Even this suggestion is problematic between studies because the number of pages will vary depending on which book is used for narration.

Future Research

Future research should examine if *D* is sensitive to the presence of an effective vocabulary intervention. Because the language samples collected at pre-test were unreliable and therefore problematic, this investigation was not able to occur in this current study. In addition to developing a standard protocol for data collection for the purposes of obtaining *D*, proper training of examiners needs to be ensured so that the time spent collecting language samples is not wasted. Training and administration protocols should be developed and documented as part

of standardizing the procedure for collecting language samples. Additionally, future research should examine if the impact of an effective vocabulary intervention is larger for children of different language abilities. Using *D* in this way, as an outcome measure, would benefit school practitioners to be able to measure *D*'s response to interventions. Lastly, future research should include longitudinal studies following the language development in this population. This could provide a better gauge of which measures are most predictive of their performance and investigate *D*'s relationship with age over time. Investigation in these areas promotes the applicability of *D* for practical uses in various education efforts.

CHAPTER 5

CONCLUSION

In conclusion, Kane's argument-based approach to validity was used to evaluate the use of *D* as a measure of lexical diversity in at-risk low-income children who are predominantly African American. The majority of the assumptions proposed in this validation study were met, with the exception of the assumption that *D* should have a linear relationship with age. This could be explained by the relatively narrow age range in this study (all children were kindergarteners). The results from this study provide support for the validity of *D. D* is not a function of sample length which provides construct validity (both full and shortened transcripts are measuring the same construct) and the significant correlation between *D* and EVT provides convergent validity. The low, but significant correlation between *D* and EVT further provides evidence that *D* measures an aspect of vocabulary that is related but different than expressive vocabulary. This supports the idea of using multiple types of measures to determine a child's language ability in order to screen for deficits and guide interventions. The evidence collected in this study provides support that *D* is generalizable across races and SES.

Perhaps one of the most valuable findings from this study comes from the comparison between African American and non-African American children. Whereas the EVT demonstrated the same discrepancies in performance between these two groups of children, there were no differences in *D*. Findings such as those reported here provide important information about the range of performance that might be expected from low-income kindergarten children. These findings suggest that *D* does not suffer from the racial-bias that standardized tests do. Because *D*

is not measured by a child knowing meanings of specific target words, it provides an unbiased and unconstrained measure of lexical diversity.

Lexical diversity measures are widespread in applied research and practice, and the findings from this study are encouraging to both children and educators. The findings from this study are encouraging to children from multicultural backgrounds whose language ability may be masked by unfamiliarity with standardized vocabulary measures. Similarly, these findings are useful to school practitioners who are interested in this specific population of children. *D* provides an alternative measure for vocabulary that can be used in addition to standardized assessments. Although somewhat time consuming, the medium to which language samples are collected are familiar to children and the conversation style allows the assessment of the active vocabulary deployed by children across different populations.

REFERENCES

Baker, S. K., Simmons, D. C., & Kaméenui, E. J. (1998). Vocabulary acquisition: Research bases. In D. C. Simmons & E. J. Kaméenui (Eds.), *What reading research tells us about children with diverse learning needs: Bases and basics* (pp. 183-217). Mahwah, NJ: Erlbaum.

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, *33*, 5-22.

Blank, M., Rose, S., & Berlin, L. (1978). PLAI: The language of learning in practice. New York: Grune & Stratton.

Brimer, M. A., & Dunn, L. (1963). *English Picture Vocabulary Test.* Windsor: NFER.

Broen, P. A. (1972). The Verbal Environment of the Language-Learning Child. *American Speech and Hearing Association Monographs*, *17*, 1-104.

Campbell, J. M., Bell, S. K., & Keith, L. K. (2001). Concurrent validity of the Peabody Picture Vocabulary Test-Third Edition as an intelligence and achievement screener for low SES African American children. *Assessment*, *8*, 85-94.

Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing bias in language assessment: Processing-dependent measures. *Journal of Speech, Language and Hearing Research*, *40*, 519-525.

Carran, D. T., & Scott, K. G. (1992). Risk assessment in preschool children: Research implications for the early detection of educational handicaps. *Topics in Early Childhood Special Education*, *12*, 196-211.

Champion, T. B., Hyter, Y. D., McCabe, A., & Bland-Stewart, L. M. (2003). A matter of vocabulary: Performances of low-income African American Head Start children on the Peabody Picture Vocabulary Test-III. *Communication Disorders Quarterly, 24*, 121-127.

Clark, E. V. (2007). Young children's uptake of new words in conversation. *Language in Society, 36*, 157-182

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.; pp. 1–74). Hillsdale, NJ: Erlbaum.

Covington, M. A. (2007). *MATTR user manual* (CASPR Research Report 2007-05). Athens, GA.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics, 17*, 94-100.

Cronbach, L. J. (1988). *Five perspectives on validity argument*. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbraum Associates, Inc.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*, 934-945.

Day, A. (1992). *Carl goes shopping*. New York, NY: Frrar, Straus and Giroux.

Day, A. (1996). *Good dog Carl*. New York, NY: Little Simon.

Dickinson, D. K., & Tabors, P. O. (Eds.). (2001). *Beginning literacy with language: Young children learning at home and school*. Baltimore, MD: Paul H. Brookes.

Dunn, L. M., & Dunn, L. M. (2007). *Peabody Picture Vocabulary Test* (4th ed.). Minneapolis, MN: Pearson Assessments.

Durán, P., Malvern, D. D., Richards, B. J., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics, 25*, 220-242.

Fagundes, D. D., Haynes, W. O., Haak, N. J., & Moran, M. J. (1998). Task variability effects on the language test performance of southern lower socioeconomic class African American and Caucasian five-year-olds. *Language, Speech, and Hearing Services in Schools*, *29*, 148-157.

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology, 22*, 397-408.

Fuller, B., Eggers-Piérola, C., Holloway, S., & Rambaud, M. (1996). Rich culture, poor markets: Why do Latino parents forego preschooling?. *Teachers College Record, 97*, 400-418.

Gardner, M. F. (1990). *Expressive One-Word Picture Vocabulary Test – Revised*. Novato, CA: Academic Therapy Publications.

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*. Hove, UK: Lawrence Erlbaum Associates, Inc.

Goodson, B., Wolf, A., Bell, S., Turner, H., and Finney, P.B. (2010). *The Effectiveness of a Program to Accelerate Vocabulary Development in Kindergarten (VOCAB).* (NCEE 2010-4014). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Paul H. Brooks.

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator*, *27*, 4-9.

Hemphill, L., & Tivnan, T. (2008). The importance of early vocabulary for literacy achievement in high-poverty schools. *Journal of Education for Students Placed at Risk, 13*, 426-451.

Hoff, E. (2006). Environmental supports for language acquisition. In D. K. Dickinson & S. B. Neuman (Eds.), *Handbook of early literacy development* (Vol. 2, pp. 163-172). New York, NY: Guilford.

Jacobson, P. F., & Walden, P. R. (2013). Lexical diversity and omission errors as predictors of language ability in the narratives of sequential Spanish-English bilinguals: A cross-language comparison. *American Journal of Speech-Language Pathology*, *22*, 554-565.

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs, 56*, 1-15.

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

Kane, M. T. (2006). The precision of measurement. *Applied Measurement in Education*, *9*, 355-379.

Kaufman, A. S., & Kaufman, N. L. (1983). Kaufman Assessment Battery for Children. Circle Pines, MN: American Guidance Services.

Kaufman, A. S. (2004). Kaufman Test of Educational Achievement-(KTEA-II). Comprehensive form. American Guidance Service, Circle Pines, MN.

Kendeou, P., van den Broek, P., White, M., & Lynch, J. S. (2007). *Comprehension in preschool and early elementary children: Skill development and strategy interventions*. In D. S. McNamara (Ed.), Reading comprehension strategies: Theories, interventions, and technologies. Mahwah, NJ: Erlbaum.

Kim, H. Y. (2013). Statistical notes for clinical researchers: assessing normal distribution using skewness and kurtosis. *Restorative Dentistry and Endodontics, 38,* 52-54.

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

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

Kresheck, J. D., & Nicolosi, L. (1973). A comparison of black and white children's scores on the Peabody Picture Vocabulary Test. *Language, Speech, and Hearing Services in Schools, 4,* 37-40.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. transcription format and programs* (3rd ed.,Vol. 1). Mahwah, NJ. Lawrence Erlbaum Associates, Inc.

Malvern, D. D., & Richards, B. J. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19,* 85-104.

Malvern, D. D., Richards, B. J., Chipere, N. and Durán, P. (2004). *Lexical diversity and language development: quantification and assessment.* New York, NY: Palgrave Macmillan.

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of Educational Research, 80,* 300-335.

McCabe, A., & Champion, T. B. (2010). A matter of vocabulary II: Low-income African American children's performance on the Expressive Vocabulary Test. *Communication Disorders Quarterly, 31,* 162-169.

McKee, G., Malvern, D. D., & Richards, B. J. (2000). Measuring vocabulary diversity using dedicated software. *Literary and linguistic computing, 15*, 323-338.

Miller, J. F. (1991). Quantifying productive language disorders. In J. F. Miller (Ed.), *Research on child language disorders: a decade of progress* (pp. 211-220). Austin, TX: Pro-Ed.

Mol, S. E., Bus, A. G., & de Jong, M. T. (2009). Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language. *Review of Educational Research, 79*, 979-1007.

Nation, K., & Snowling, M. J. (2004). Beyond phonological skills: Broader language skills contribute to the development of reading. *Journal of Research in Reading, 27*, 342-356.

National Assessment Governing Board (2012). Reading Framework for the 2013 National Assessment of Educational Progress. Washington, DC: Author.

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

Neuman, S. B. (2009). *Changing the odds for children at risk*. New York, NY: Teachers College Press.

Neuman, S. B., & Celano, D. (2001). Access to print in low- and middle-income communities: An ecological study of four neighborhoods. *Reading Research Quarterly, 36*, 8-26.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Owen, A. J., & Leonard, L. B. (2002). Lexical diversity in the spontaneous speech of children with specific language impairment: Application of D. *Journal of Speech, Language and Hearing Research, 45*, 927-937.

Qi, C. H., Kaiser, A. P., Milan, S., & Hancock, T. (2006). Language performance of low-income African American and European American preschool children on the PPVT-III. *Language, Speech, and Hearing Services in Schools*, *37*, 5-16.

Restrepo, M. A., Schwanenflugel, P. J., Blake, J., Neuharth-Pritchett, S., Cramer, S. E., & Ruston, H. P. (2006). Performance on the PPVT-III and the EVT: Applicability of the measures with African American and European American preschool children. *Language, Speech, and Hearing Services in Schools*, *37*, 17-27.

Richards, B. J., & Malvern, D. D. (1997). *Quantifying lexical diversity in the study of language development.* Reading: New Bulmershe Papers, University of Reading.

Ricciuti, A.E., St. Pierre, R.G., Lee, W., Parsad, A. & Rimdzius, T. (2004) Third national Even Start evaluation: Follow-up findings from the experimental design study. U. S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Washington, DC: 2004.

Ruston, H. P. (2007). Effects of a conversation intervention on expressive vocabulary of young Children (Unpublished doctoral dissertation). University of Georgia, Athens, Ga.

Ruston, H. P., & Schwanenflugel, P. J. (2010). Effects of a conversation intervention on the expressive vocabulary development of prekindergarten children. *Language, speech, and hearing services in schools, 41*, 303-313.

Rvachew, S., Ohberg, A., Grawburg, M, & Heyding, J. (2003). Phonological awareness and phonemic perception in 4-year-old children with delayed expressive phonology skills. *American Journal of Speech-Language Pathology*, *12*, 463-471.

Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), Specific reading disability: A view of the spectrum (pp. 75–119). Timonium, MD: York Press.

Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology, 57*, 1-10.

Schwanenflugel, P. J., Hamilton, C. E., Bradley, B. A., Ruston, H. P., Neuharth-Pritchett, S., & Restrepo, M. A. (2005). Classroom practices for vocabulary enhancement in prekindergarten: Lessons from PAVEd for success. In E. H. Hiebert & M. Kamil (Eds.) *Bringing scientific research to practice: Vocabulary* (pp. 155-177). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals–Fourth Edition*. San Antonio, TX: The Psychological Corporation.

Sénéchal, M., & LeFevre, J. A. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development*, *73*, 445-460.

Seymore, H. N., Roeper, T. W., deVilliers, J., deVilliers, P. A. (2009). *Diagnostic Evaluation of Language Variation (DELV–Screening Test).* San Antonio, TX: Pearson Education.

Silverman, S., & Ratner, N. B. (2002). Measuring lexical diversity in children who stutter: Application of *vocd*. *Journal of Fluency Disorder*, *27*, 289-304.

Spira, E. G., Bracken, S. S., & Fischel, J. E. (2005). Predicting improvement after first-grade reading difficulties: the effects of oral language, emergent literacy, and behavior skills. *Developmental Psychology, 41*, 225-234.

Stahl, S. A. (1999). *Vocabulary Development (From Reading Research to Practice, V. 2).* Cambridge, MA: Brookline Books.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*, 360-407.

Sternberg, R. J. (Ed.). (1988). *The nature of creativity: Contemporary psychological perspectives*. New York, NY: Cambridge University Press.

Stockman, I. J. (2000). The new Peabody Picture Vocabulary Test-III: An illusion of unbiased assessment?. *Language, Speech, and Hearing Services in Schools, 31*, 340-353.

Thomas-Tate, S., Washington, J., Craig, H., & Packard, M. (2006). Performance of African American preschool and kindergarten students on the Expressive Vocabulary Test. *Language, Speech, and Hearing Services in Schools, 37*, 143-149.

Thordardottir, E. T., Ellis Weismer, S. (2001). High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment. *International Journal of Language & Communication Disorders, 36*, 221-244.

U. S. Census Bureau (March 2001). Current Population Survey. *Census.gov.* Retrieved April 10, 2014, from http://www.census.gov.

U. S. Census Bureau (2008). Poverty. Retrieved June 7, 2014, from http://www.census.gov/hhes/www/poverty/

U.S. Department of Education (2009, March 3). Early Reading First. *Ed*.gov. Retrieved November 2, 2013, from http://www.ed.gov.

U.S. Department of Health and Human Services (U.S. DHHS, 2006). FACES 2003 Research Brief: Children's Outcomes and Program Quality in Head Start. Retrieved November 2, 2013, from http://www.acf.hhs.gov/

Washington, J. A., & Craig, H. K. (1992). Performances of low-income, African American preschool and kindergarten children on the Peabody Picture Vocabulary Test-Revised. *Language, Speech, and Hearing Services in Schools, 23,* 329-333.

Washington, J. A., & Craig, H. K. (1999). Performances of at-risk, African American preschoolers on the Peabody Picture Vocabulary Test-III. *Language, Speech, and Hearing Services in Schools, 30,* 75-82.

Washington, J. A., & Craig, H. K. (2004). A language screening protocol for use with young African American children in urban settings. *American Journal of Speech-Language Pathology, 13,* 329-340.

Webb, M. Y. L., Cohen, A. S., & Schwanenflugel, P. J. (2008). Latent class analysis of differential item functioning on the Peabody Picture Vocabulary Test–III. *Educational and Psychological Measurement, 68,* 335-351.

White, T. G., Graves, M. F., & Slater, W. H. (1990). Growth of reading vocabulary in diverse elementary schools: Decoding and word meaning. *Journal of Educational Psychology, 82,* 281-290.

Williams, K. T. (1997). *Expressive Vocabulary Test*. Circle Pines, MN: American Guidance Service.

Williams, K. T. (2007). *Expressive Vocabulary Test, Second Edition*. Circle Pines, MN: American Guidance Service.

Wright, H. H., Silverman, S., & Newhoff, M. (2003) Measures of lexical diversity in aphasia. *Aphasiology. 17,* 443-452.

APPENDIX A

SIGNED RESTRICTED-USE IES DATA LICENSE

LICENSE FOR THE USE OF
INDIVIDUALLY IDENTIFIABLE INFORMATION
PROTECTED UNDER
THE EDUCATION SCIENCES REFORM ACT OF 2002
AND PROTECTED, AS APPLICABLE, UNDER
THE E-GOVERNMENT ACT OF 2002, TITLE V
THE CONFIDENTIAL INFORMATION PROTECTION AND STATISTICAL EFFICIENCY ACT OF 2002,
THE FAMILY EDUCATIONAL RIGHTS AND PRIVACY ACT
AND THE PRIVACY ACT OF 1974

WHEREAS, the Institute of Education Sciences (IES) of the United States Department of Education has collected and maintains individually identifiable information, the confidentiality of which is protected by section 183 of the Education Sciences Reform Act of 2002 (ESRA) (PL 107-279) (20 U.S.C. 9573), and, as applicable, by the Privacy Act of 1974 (5 U.S.C. 552a); the Family Educational Rights and Privacy Act (FERPA) (20 U.S.C. 1232g); and Title V, subtitle A of the E-Government Act of 2002 (CIPSEA) (PL 107-347) (44 U.S.C. 3501 note); and

WHEREAS, IES wishes to make the data available for statistical, research, or evaluation purposes to requestors qualified and capable of research and analysis consistent with the statistical, research, or evaluation purposes for which the data were provided or are maintained, but only if the data are used and protected in accordance with the terms and conditions stated in this license (License), upon receipt of such assurance of qualification and capability, it is hereby agreed between

University of Georgia,
(Insert the name of the agency or organization to be licensed)

hereinafter referred to as the "Licensee", and IES that:

I. INFORMATION SUBJECT TO THIS AGREEMENT

A. All data containing individually identifiable information about students, their families, and their schools maintained by IES under section 183 of the Education Sciences Reform Act of 2002, that are provided to the Licensee and all information derived from those data, and all data resulting from merges, matches, or other uses of the data provided by IES with other data are subject to this License and are referred to in this License as subject data.

B. Subject data under this License may be in the form of CD-ROMs, electronic data, hard copy, etc. The Licensee may only use the subject data in a manner and to a purpose consistent with:

1. The statistical, research, or evaluation purpose for which the data are maintained. All subject data that include individually identifiable information are protected under the Privacy Act, ESRA, and/or

1

CIPSEA and may be used only for statistical, research, or evaluation purposes consistent with purposes for which the data were collected and /or are maintained (Licensee's description of the research and analysis which is planned is attached and made a part of this License - Attachment No. 1.);

2. Subject data that includes personally identifiable information from students' education records are protected under FERPA and may only be used for the evaluation of Federally-supported education programs or for conducting studies, for, or on behalf of, educational agencies or institutions to improve instruction. (Licensee's description of the evaluation or study which is planned is attached and made a part of this License - Attachment No. 1.);

3. The limitations imposed under the provisions of this License; and

4. Section 183 of the Education Sciences Reform Act of 2002 (20 U.S.C. 9573); and, as applicable, Title V, subtitle A of the E-Government Act of 2002 (44 U.S.C. 3501 note); the Privacy Act of 1974 (5 U.S.C. 552a), and the Family Educational Rights Protection Act (20 U.S.C. 1232g) which are attached to and made a part of this License (Attachment No. 2.)

II. INDIVIDUALS WHO MAY HAVE ACCESS TO SUBJECT DATA

A. There are four categories of individuals that the Licensee may authorize to have access to subject data. The four categories of individuals are as follows:

1. The Principal Project Officer (PPO) is the most senior officer in charge of the day-to-day operations involving the use of subject data and is responsible for liaison with IES.

2. Professional/Technical staff (P/T) conduct the research for which this License was issued.

3. Support staff includes secretaries, typists, computer technicians, messengers, etc. Licensee may disclose subject data to support staff who come in contact with the subject data in course of their duties only to the extent necessary to support the research under this License.

4. The System Security Officer (SSO) is responsible for maintaining the day-to-day security of the licensed data, including the implementation, maintenance, and periodic update of the Security Plan to protect the data in strict compliance with statutory and regulatory requirements.

B. Licensee may disclose subject data to only to only seven (7) staff, including the PPO, SSO, P/TS, and support staff, unless IES provides written authorization for a larger number of P/TS.

III. LIMITATIONS ON DISCLOSURE

A. Licensee shall not use or disclose subject data for any administrative purposes nor may the subject data be applied in any manner to change the status, condition, or public perception of any individual regarding whom subject data is maintained. (Note: Federal Law pre-empts any State law that might require the reporting or dissemination of these data for any purpose other than the statistical, research, and evaluation purposes for which they were collected and/or are maintained.)

B. Licensee shall not disclose subject data or other information containing, or derived from, subject data at fine levels of geography, such as school district, institution, or school, to anyone other than IES employees working in the course of their employment or individuals for whom access is authorized under this License agreement. Licensee may make disclosures of subject data to individuals other than those specified in this License only if those individuals have executed an Affidavit of Nondisclosure and the Licensee has obtained advance written approval from the IES Data Security Office.

C. Licensee shall not make any publication or other release of subject data listing information regarding individuals or specific educational institutions even if the individual respondent identifiers have been removed.

D. Licensee may publish the results, analysis, or other information developed as a result of any research based on subject data made available under this License only in summary or statistical form so that the identity of individuals or specific educational institutions contained in the subject data is not revealed.

IV. ADMINISTRATIVE REQUIREMENTS

A. The research conducted under this License and the disclosure of subject data needed for that research must be consistent with the statistical, research, or evaluation purpose for which the data were supplied. The subject data may not be used to identify individuals or specific educational institutions for recontacting unless Licensee has obtained advance written approval from the IES Data Security Office.

B. Execution of Affidavits of Nondisclosure.

1. Licensee shall provide a copy of this agreement, together with the Security Plan (Attachment No. 3) to the SSO and to each P/T and support staff person of the Licensee who will have access to subject data and shall require each of those individuals to execute an Affidavit of Nondisclosure (Attachment No. 4).

2. The Licensee must ensure that each individual who executes an Affidavit of Nondisclosure reads and understands the materials provided to her or him before executing the Affidavit.

3. Licensee shall ensure that each Affidavit of Nondisclosure is notarized upon execution.

4. Licensee may not permit any individual specified in paragraph II.A. to have access to subject data until the procedures in paragraphs IV.B.1. through 3 of this License are fulfilled for that individual.

5. Licensee shall promptly, after the execution of each Affidavit, send the original Affidavit to the IES Data Security Office and shall maintain a copy of each Affidavit at the Licensee's secured facility protected under this License.

C. Notification regarding authorized individuals to IES.

1. Licensee shall promptly notify the IES Data Security Office when the SSO, or any P/T or support staff who has been authorized to have access to subject data no longer has access to those data.

D. Publications made available to IES.

1. Licensee shall provide the IES Data security Office a copy of each publication containing information based on subject data or other data product based on subject data before they are made available to individuals who have not executed an Affidavit of Nondisclosure.

2. Because the publication or other release of research results could raise reasonable questions regarding disclosure of individually identifiable information contained in subject data, copies of the proposed publication or release must be provided to the IES Data Security Office before that disclosure is made so that IES may advise whether the disclosure is authorized under this License and the provisions of section 183 of the Education Sciences Reform Act of 2002; Title V, subtitle A of the E-Government Act of 2002; the Privacy Act of 1974; and the Family Educational Rights and Privacy Act. Licensee agrees not to publish or otherwise release research results provided to IES if IES advises that such disclosure is not authorized.

E. Licensee shall notify the IES Data Security Office immediately upon receipt of any legal, investigatory, or other demand for disclosure of subject data.

F. Licensee shall notify the IES Data Security Office immediately upon discovering any breach or suspected breach of security or any disclosure of subject data to unauthorized parties or agencies.

G. Licensee agrees that representatives of IES have the right to make unannounced and unscheduled inspections of the Licensee's facilities, including any associated computer center, to evaluate compliance with the terms of this License and the requirements of section 183 of the Education Sciences Reform Act of 2002; Title V, subtitle A of the E-Government Act of 2002; the Privacy Act of 1974; and the Family Educational Rights and Privacy Act.

V. SECURITY REQUIREMENTS

A. Maintenance of, and access to, subject data.

1. Licensee shall retain the original version of the subject data at a single location and may make no copy or extract of the subject data available to anyone except the SSO or a P/T staff member as necessary for the purpose of the statistical research for which the subject data were made available to the Licensee.
2. Licensee shall maintain subject data (whether maintained on a personal computer or on printed or other material) in a space that is limited to access by the PPO, SSO, and authorized P/T staff.
3. Licensee shall ensure that access to subject data maintained in computer memory is controlled by password protection. Licensee shall maintain all print-outs, CD-ROMS, personal computers with subject data on hard disks, or other physical products containing individually identifiable information derived from subject data in locked cabinets, file drawers, or other secure locations when not in use.
4. Licensee shall ensure that all printouts, tabulations, and reports are edited for any possible disclosures of subject data.
5. Licensee shall establish security procedures to ensure that subject data cannot be used or taken by unauthorized individuals.
6. Licensee shall not permit removal of any subject data from the limited access space protected under the provisions of this License as required in the attached Security Plan (Attachment No. 3.), without first notifying, and obtaining written approval from, IES.

B. Retention of subject data.

Licensee shall return to the IES Data Security Office all subject data, or destroy those data under IES supervision or by approved IES procedures when the statistical analysis, research, or evaluation that is the subject of this agreement has been completed or this License terminates, whichever occurs first. Licensee, as part of its responsibilities discussed herein, agrees to submit a completed Close-out Certification Form to the IES Data Security Office.

C. Compliance with established security procedures.

Licensee shall comply with the security procedures described in the Security Plan (Attachment No. 3 to this License).

VI. PENALTIES

A. Any violation of the terms and conditions of this License may subject the Licensee to immediate revocation of the License by IES.

1. The IES official responsible for liaison with the Licensee shall initiate revocation of this License by written notice to Licensee indicating the factual basis and grounds for revocation.

2. Upon receipt of the notice specified in paragraph VI.A.1 of this License, the Licensee has thirty (30) days to submit written argument and evidence to the Director of IES indicating why the License should not be revoked.

3. The Director of IES shall decide whether to revoke the License based solely on the information contained in the notice to the Licensee and the Licensee's response and shall provide written notice of the decision to the Licensee within forty-five (45) days after receipt of Licensee's response. The Director of IES may extend this time period for good cause.

B. Any violation of this License may also be a violation of Federal criminal law under the Privacy Act of 1974 (5 U.S.C. 552a(i)); section 183 of the Education Sciences Reform Act of 2002 (20 U.S.C. 9573(d)(2); and/or Title V, subtitle A of the E-Government Act of 2002. Alleged violations under section 183 of the Education Sciences Reform Act of 2002 and Title V, subtitle A of the E-Government Act of 2002 are subject to prosecution by the Offices of the United States Attorney. The penalty for violation of section 183 of the Education Sciences Reform Act of 2002 and Title V, subtitle A of the E-Government Act of 2002, is a fine of not more than $250,000 and imprisonment for a period of not more than five years.

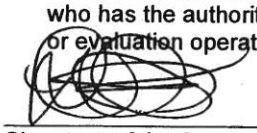
VII. PROCESSING OF THIS LICENSE

A. The term of this License shall be for 2 years. If, before the expiration of this License, the Director of IES establishes regulatory standards for the issuance and content of Licenses, the Licensee agrees to comply with the regulatory standards.

B. This License may be amended, extended, or terminated by mutual written agreement between the Licensee and the Director of IES. Any amendment must be signed by a Senior Official specified in paragraph VII.C. of this License, PPO, and the Director of IES and is effective on the date that all required parties have signed the amendment.

C. The Senior Official (SO), who cannot be the same individual designated as the PPO, having the legal authority to bind the organization to the terms of the License, shall sign this License below. The SO certifies, by his/her signature, that -

1. The organization has the authority to undertake the commitments in this License;
2. The SO has the legal authority to bind the organization to the provisions of this License; and
3. The PPO is the most senior subject matter officer for the Licensee who has the authority to manage the day-to-day statistical, research, or evaluation operations of the Licensee.

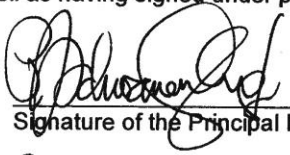


Signature of the Senior Official 6/26/13 Date

Regina Smith
Type/Print Name of Senior Official

Title: Associate Vice President of Research Telephone: (706) 583-0443

D. The individual described in paragraph II.A1. as the PPO shall sign this License below. If the SO also acts as the chief statistical officer for the Licensee; viz. as the PPO, the SO shall likewise sign under this paragraph as well as having signed under paragraph VII.C.



Signature of the Principal Project Officer 6/26/13 Date

Paula Schwanenflugel
Type/Print Name of the Principal Project Officer

Title: Professor Telephone: (706) 542-4273

E. The Director of the Institute of Education Sciences or Designee issues this License to

University of Georgia. The License is effective as of the date of the Director of IES or Designee's signature below, or such other period specified in the Licensee's request for the License.

Signature of Director of IES or Designee

Commissioner, NCES
Title

Jack Buckley
Type/Print Name of Director of IES or Designee

7/24/2013
Date

IES License Control Number: 13030004

APPENDIX B

SIGNED IES NOTARIZED AFFIDAVITS

Affidavit of Nondisclosure

Graduate Student
(Job Title)

7/13/2013
(Date Assigned to Work with IES Data)

University of Georgia
(Organization, State or Local Agency Name)

All NCES Data
(NCES Database or File Containing Individually Identifiable Information*)

(Organization or Agency Address)

I, Stephanie Lai, do solemnly swear (or affirm) that when given access to the subject IES database or file, I will not -

(i) use or reveal any individually identifiable information furnished, acquired, retrieved or assembled by me or others, under the provisions of Section 183 of the Education Sciences Reform Act of 2002 (P.L. 107-279) for any purpose other than statistical, research, or evaluation purposes specified in the IES survey, project or contract;

(ii) make any disclosure or publication whereby a sample unit or survey respondent (including students and schools) could be identified or the data furnished by or related to any particular person or school under these sections could be identified; or

(iii) permit anyone other than the individuals authorized by the Director of the Institute of Education Sciences to examine the individual reports.

(Signature)

[The penalty for unlawful disclosure is a fine of not more than $250,000 (under 18 U.S.C. 3571) or imprisonment for not more than five years (under 18 U.S.C. 3559), or both. The word "swear" should be stricken out when a person elects to affirm the affidavit rather than to swear to it.]

City/County of Clarke Commonwealth/State of Georgia.
Sworn to and subscribed before me this 1st day of April, 2013. Witness my hand and official Seal.

(Notary Public/Seal)

My commission expires 9/20/14.

* Request all subsequent follow-up data that may be needed. This form cannot be amended by NCES, so access to databases not listed will require submitting additional notarized Affidavits.

Form last revised 09/27/12

Affidavit of Nondisclosure

Full Professor
(Job Title)

University of Georgia
(Organization, State or Local Agency Name)

(Organization or Agency Address)

7/16/2013
(Date Assigned to Work with IES Data)

All NCES/IES Data
(NCES Database or File Containing Individually Identifiable Information*)

I, Paula Schwanenflugel, do solemnly swear (or affirm) that when given access to the subject IES database or file, I will not -

(i) use or reveal any individually identifiable information furnished, acquired, retrieved or assembled by me or others, under the provisions of Section 183 of the Education Sciences Reform Act of 2002 (P.L. 107-279) for any purpose other than statistical, research, or evaluation purposes specified in the IES survey, project or contract;

(ii) make any disclosure or publication whereby a sample unit or survey respondent (including students and schools) could be identified or the data furnished by or related to any particular person or school under these sections could be identified; or

(iii) permit anyone other than the individuals authorized by the Director of the Institute of Education Sciences to examine the individual reports.

(Signature)

[The penalty for unlawful disclosure is a fine of not more than $250,000 (under 18 U.S.C. 3571) or imprisonment for not more than five years (under 18 U.S.C. 3559), or both. The word "swear" should be stricken out when a person elects to affirm the affidavit rather than to swear to it.]

City/County of Gwinnett Commonwealth/State of GA.
Sworn to and subscribed before me this 7th day of March 2013. Witness my hand and official Seal.

(Notary Public/Seal)

My commission expires Nov 5, 2014

* Request all subsequent follow-up data that may be needed. This form cannot be amended by NCES, so access to databases not listed will require submitting additional notarized Affidavits.

Form last revised 09/27/12

APPENDIX C

SIGNED IES SECURITY PLAN FORM

Security Plan Form

Institute of Education Sciences (IES)
Restricted-use Data

Name of Institution / Organization: University of Georgia

PPO Name: Paula Schwanenflugel

PPO Address:
(no P.O. Box number; specify building name, department, and room number)

(Provide street address, city, state, zip code, department and building name, and office/room number.)
110 Carlton Street
Aderhold Hall
University of Georgia
Department of Educational Psychology and Instructional Technology
Room 325R
Athens, GA 30605

PPO Phone Number: 706-542-4273

Type of Security Plan: New ☑ Renewal ☐ Modification ☐

License Number: 13030004

Physical Location of Data

Project Office Address:
(no P.O. Box number; specify building name, department, and room number)

(Provide street address, city, state, zip code, department and building name, and office/room number.)
110 Carlton Street
Aderhold Hall
University of Georgia
Department of Educational Psychology and Instructional Technology
Room 320K
Athens, GA 30605

Project Office Phone Number: 706-296-6317

***Note:** The restricted-use data and computer must be secured and used **only** at this location. When the data are not being used, the data must be stored under lock and key at this location. Only authorized users of the data, as listed on the License, may have key access to this secure project office/room.*

Physical Security of Data

Describe Building Security:
(Describe building security arrangements where project office is located.)

The project office is located in Aderhold Hall at the University of Georgia. The building is monitored through the Andover Continuum facilities management program. The hours between 7:00 AM and 8:00 PM are considered normal hours of operation. The hours between 8:01 PM and 6:59 AM are considered after hours, and the building is locked with only activated University of Georgia cards granted access to the building. During after hours, exterior doors are locked and the building is monitored by the University of Georgia Police Department. If anything goes wrong (e.g., doors are propped open or a door lock malfunctions) an alarm system is set off, and the University's police department is alerted. Thereafter, officers are sent to investigate the building.

Describe Project Office Security:
(Describe project office security arrangements for the room where the computer and data will be located.)

The project office is located in Aderhold Hall at the University of Georgia. The project office will be locked, with keys being distributed to only those listed on the formal request form. No other personnel will be given access to the project office.

Computer Security Requirements

Describe Computer System:
(Please read the Note below. Computer security must follow the requirements listed below.)

The Dell computer is a standalone, desktop computer that is operated using Windows 7.

Computer Operating System: Windows 7

Anti-Virus Software Installed on Computer: Trend Micro

Note: *The restricted-use data must be copied to and run on a standalone, desktop computer.* ***Use of a laptop computer, external hard drive, or USB memory stick is strictly prohibited.*** *Absolutely no restricted-use data may be copied onto a server or computer that is attached to a modem or network (LAN) connection. Prior to attaching the computer to a modem or LAN connection, the restricted-use data must be purged and overwritten on the computer.*

The following physical location and computer security procedures must be implemented when in possession of restricted-use data. By checking the box next to each security procedure, you signify that these security procedures will be implemented for the duration of the project and License period:

- Only authorized users listed on the License will have access to the secure room. Access will be limited to the secure room/project office by locking the office when away from the office. ☑
- Data will only be secured, accessed and used within the secure project office/room (as specified on page 1 of this plan). ☑
- A password will be required as part of the computer login process. ☑
- The password for computer access will be unique and contain 6 to 8 characters with at least one non-alphanumeric character. ☑

- The computer password will change at least every 3 months or when project staff leave. ☑
- Read-only access will be initiated for the original data. ☑
- An automatic password protected screensaver will enable after 5 minutes of inactivity. ☑
- No routine backups of the restricted-use data will be made. ☑
- Project office room keys will be returned and computer login will be disable within 24 hours after any user leaves the project. The PPO will notify IES of staff changes. ☑
- Restricted-use data will **not** be placed on a server (network), laptop computer, USB memory stick, or external hard drive. ☑
- The data will be removed from the project computer and overwritten, whether at the end of the project or when reattaching a modem or LAN connection. ☑
- Post Warning notification: During the computer log-in process, a warning statement (shown below) will appear on the computer screen before access is granted. If it is not possible to have the warning appear on the screen, it must be typed and attached to the computer monitor in a prominent location. ☑

WARNING

U.S. Government Restricted-use Data

Unauthorized Access to Data (Individually Identifiable Information) on this Computer is a Violation of Federal Law and will Result in Prosecution.

Do You Wish to Continue? (Y)es or (N)o

NOTICE

Proposed Publications Using Restricted-use Data

Sample Surveys and Evaluations

Licensees are required to round all unweighted sample size numbers to the nearest ten (nearest 50 for the Early Childhood Longitudinal Study Birth Cohort) in all information products (i.e., proposals, presentations, papers or other documents that are based on or use restricted-use data). Licensees are required to provide a draft copy of each information product that is based on or uses restricted-use data to the IES Data Security Office for a disclosure review. In the case of information products that are based on or use FERPA-protected restricted use data, the IES Data Security Office will also review the product to determine if, consistent with the approved project proposal, the Licensee used the data to conduct a study to improve instruction or as an "authorized representative of the Secretary" to evaluate a Federally supported education program. The Licensee must not release the information product to any person not authorized to access the data you are using until formally notified by IES that no potential disclosures were found and, if applicable, that no FERPA issues were identified. This review process usually takes 3 to 5 business days.

The PPO shall also forward a final copy of any public presentations or reports published or released that are based on or use restricted-use to the IES Data Security Office to provide feedback on uses of ESRA data.

Administrative Record/Universe Data

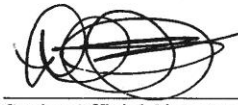
Licensees are required to follow the disclosure avoidance procedures transmitted with the restricted-use data in all information products (i.e., proposals, presentations, papers or other documents that are based on or use restricted-use data). Licensees are required to provide a draft copy of each information product that is based on or uses restricted-use data to the IES Data Security Office for a disclosure review. In the case of information products that are based on or use FERPA-protected restricted-use data, the IES Data Security Office will also review the product to determine if, consistent with the approved project proposal, the Licensee used the data to conduct a study to improve instruction or as an "authorized representative of the Secretary" to evaluate a Federally supported education program. The Licensee must not release the information product to any person not authorized to access the data you are using until formally notified by IES that no potential disclosures were found and, if applicable, that no FERPA issues were identified. This review process usually takes 3 to 5 business days.

The PPO shall also forward a final copy of any public presentations or reports published or released that are based on, or use FERPA-protected restricted-use data to the IES Data Security Office.

Signature Page – Management Review and Approval

I have reviewed the requirements of the License agreement and the security procedures in this plan that describe the required protection procedures for securing, accessing and using the restricted-use data.

I hereby certify that the computer system, physical location security procedures, and access procedures meet all of the License requirements and will be implemented for the duration of the project and License period.



Regina A. Smith, Ph.D.
Associate Vice President for Research

Senior Official Signature

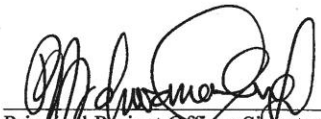
Date

Regina Smith

Senior Official Name & Title (print)

706-583-0443

Phone Number



Principal Project Officer Signature

6/26/13

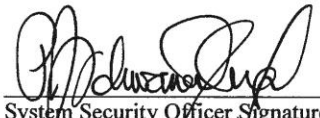
Date

Paula Schwanenflugel

Principal Project Officer Name & Title (print)

706-542-4243

Phone Number



System Security Officer Signature

6/26/13

Date

Paula Schwanenflugel

System Security Officer Name & Title (print)

706-542-4243

Phone Number

Note: The National Center for Education Statistics (NCES) processes licenses and disseminates restricted-use data for all centers in the Institute of Education Sciences (IES) including the National Center for Education Research (NCER), the National Center for Education Statistics (NCES), the National Center for Education Evaluation (NCEE), and the National Center for Special Education Research (NCSER).

APPENDIX D

LANGUAGE SAMPLE TRANSCRIBED IN CHAT SPECIFICATIONS

@Begin
@Languages: en
@Participants: CHI Sample Child, INT Investigator
@ID: en.sample=CHI
*INT: okay what is happening in the picture.
*CHI: the dog going to help them watch the baby.
*INT: okay what is happening in this picture.
*CHI: the baby is pulling on the dog ear.
*INT: okay.
*CHI: the baby is on the dog.
*INT: okay what are they doing what else is happening.
*CHI: they fixing to go into an elevator.
*INT: okay.
*CHI: they <they>[/] went up the elevator they went down to a xxx he put a baby in a truck and he rained on it he fell in the toys.
*INT: what is happening in this picture?
*CHI: the dog is holding him by biting his shirt and his sister is looking at him.
*INT: okay.
*CHI: he give the baby the glove and they went somewhere with the hat and the glove.
*INT: okay.
*CHI: they took some pictures so it could be on the television he <he he he>[/] wanted to go in the living room he <he he>[/] had on the couch.
*INT: okay.
*CHI: they went to the store.
*INT: okay.
*CHI: to <to>[/] get some food then they ate it.
*INT: okay what else what is happening here.
*CHI: they went to the pet store they <they>[/] buy they got a pet.
*INT: what happened right here.
*CHI: he is on a dog the dog let go of the pets.
*INT: what else happened here.
*CHI: they trying to put them back it is out the cage they went to the <the the>[/] malls they bought a they go a carpet and c.
*INT: can you tell me anything else happening in this picture.
*CHI: no.
*INT: okay.
*CHI: mom was about to come home the dog put the baby back in the blue basket mom came down stairs the baby was in the basket every kind of footprints the end.
*INT: this is the end of sample identification number is sample.
@End

APPENDIX E

FREQ COMMAND AND OUTPUT

```
> freq +t*CHI +r6 sample.cha
freq sample.cha
freq sample.cha
Thu Apr 10 12:28:27 2014
freq (10-Apr-2014) is conducting analyses on:
ONLY speaker main tiers matching: *CHI;

 18 a
  1 about
  7 and
  2 at
  1 ate
  7 baby
  2 back
  2 basket
  2 be
  1 beside
  1 biting
  1 blue
  1 book
  1 bought
  1 but
  1 buy
  1 by
  1 c
  1 cage
  1 came
 11 can
  1 carpet
  1 cavity
  1 christmas
  2 come
  1 couch
  1 could
  8 dog
  2 down
  1 ear
  2 elevator
  1 end
  1 every
  1 everything
  1 fell
```

1 fix
2 fixing
1 flower
2 fly
1 food
1 footprints
2 found
3 get
3 give
2 glove
1 gloves
4 go
1 going
1 got
1 had
2 has
1 hat
1 have
25 he
1 help
5 him
2 his
1 holding
1 home
1 i
1 if
5 in
1 into
7 is
8 it
1 kind
1 let
2 library
1 like
1 living
1 looking
1 malls
2 mom
1 night
5 no
2 of
6 on
1 one
1 only
1 or
2 out

2 pet
1 pets
3 pick
1 pictures
1 plant
2 play
1 presents
1 pulling
1 pushing
4 put
1 rained
1 reindeer
2 room
1 see
1 shirt
1 sister
1 sit
1 sleigh
1 so
3 some
1 somewhere
1 stairs
2 store
1 stuff
1 supposed
1 teeth
1 television
2 tell
2 thank
4 that
30 the
2 them
1 then
22 they
1 things
17 to
1 took
1 tooth
2 toys
1 tree
1 truck
1 trying
1 under
1 up
1 want
1 wanted

  2 was
  1 watch
  8 went
  1 what
  1 whenever
  1 will
  3 with
 10 yes
 12 you
  2 your
---
  137 Total number of different item types used
  376 Total number of items (tokens)
0.364 Type/Token ratio

APPENDIX F

VOCD COMMAND AND OUTPUT

> vocd +t*CHI +r6 sample.cha
vocd +t*CHI +r6 sample.cha
Thu Apr 10 12:28:27 2014
vocd (10-Apr-2014) is conducting analyses on:
  ONLY speaker main tiers matching: *CHI;

yes he give you presents
yes with things that you like
he has a sleigh
it can fly
yes
he has a reindeer that can fly
yes
yes he you can tell him what that you want at night he will put it under your christmas tree
no
he can fix your teeth whenever he can give you stuff
yes
he can get you toys
yes but it is only supposed to be one
yes
he can tell you if you have a cavity
yes or a tooth fixing to come out
yes
no
play
you can sit beside him
you can play with him
no
a dog pushing a plant i see a flower
the dog going to help them watch the baby
the baby is pulling on the dog ear
the baby is on the dog
they fixing to go into a elevator
they went up the elevator they went down to a room he put a baby in a truck and he rained on it
he fell in the toys
the dog is holding him by biting his shirt and his sister is looking at him
thank you
no
they went to library to pick a book they found some gloves and everything
thank you he give the baby the glove and they went somewhere with the hat and the glove
they took some pictures so it could be on the television he wanted to go in the living room he had on the couch

they went to the store
to get some food then they ate it
they went to the pet store they buy they got a pet
he
he is on a dog the dog let go of the pets
they trying to put them back it is out the cage they went to the malls they bought a they go a carpet and c
no
mom was about to come home the dog put the baby back in the blue basket mom came down stairs the baby was in the basket every kind of footprints the end

tokens	samples	ttr	st.dev	D
35	100	0.7686	0.064	44.667
36	100	0.7628	0.062	44.148
37	100	0.7673	0.068	46.806
38	100	0.7603	0.057	45.809
39	100	0.7500	0.063	43.875
40	100	0.7375	0.061	41.440
41	100	0.7434	0.062	44.155
42	100	0.7400	0.059	44.229
43	100	0.7281	0.060	41.930
44	100	0.7257	0.057	42.234
45	100	0.7171	0.061	40.902
46	100	0.7178	0.059	42.000
47	100	0.7255	0.056	45.070
48	100	0.7229	0.048	45.267
49	100	0.7092	0.053	42.371
50	100	0.7096	0.051	43.348

D: average = 43.641; std dev. = 1.640
D_optimum <43.56; min least sq val = 0.001>

tokens	samples	ttr	st.dev	D
35	100	0.7629	0.063	42.945
36	100	0.7508	0.062	40.726
37	100	0.7632	0.060	45.519
38	100	0.7500	0.064	42.750
39	100	0.7577	0.064	46.201
40	100	0.7448	0.064	43.460
41	100	0.7568	0.058	48.288
42	100	0.7317	0.062	41.896
43	100	0.7288	0.059	42.118
44	100	0.7320	0.056	43.998
45	100	0.7240	0.063	42.732
46	100	0.7304	0.054	45.523
47	100	0.7091	0.066	40.632

48	100	0.7138	0.060	42.713
49	100	0.7016	0.057	40.424
50	100	0.7054	0.053	42.226

D: average = 43.259; std dev. = 2.111
D_optimum <43.17; min least sq val = 0.001>

tokens	samples	ttr	st.dev	D
35	100	0.7726	0.060	45.927
36	100	0.7611	0.063	43.649
37	100	0.7616	0.061	45.018
38	100	0.7605	0.064	45.891
39	100	0.7531	0.067	44.787
40	100	0.7570	0.054	47.165
41	100	0.7368	0.063	42.291
42	100	0.7467	0.071	46.215
43	100	0.7388	0.063	44.939
44	100	0.7295	0.069	43.295
45	100	0.7300	0.060	44.408
46	100	0.7204	0.057	42.701
47	100	0.7217	0.064	43.982
48	100	0.7296	0.054	47.242
49	100	0.7151	0.062	43.976
50	100	0.7116	0.057	43.895

D: average = 44.711; std dev. = 1.427
D_optimum <44.66; min least sq val = 0.000>

VOCD RESULTS SUMMARY
==========
Types,Tokens,TTR: <137,344,0.398256>
D_optimum values: <43.56, 43.17, 44.66>
D_optimum average: 43.80