DIGITAL COMMONS
@ UNIVERSITY OF SOUTH FLORIDA

University of South Florida
Digital Commons @ University of South Florida

USF Tampa Graduate Theses and Dissertations

USF Graduate Theses and Dissertations

March 2022

Effective Statistical and Machine Learning Methods to Analyze Children's Vocabulary Learning

Houston T. Sanders
University of South Florida

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

 Part of the Education Commons, and the Statistics and Probability Commons

Scholar Commons Citation

Sanders, Houston T., "Effective Statistical and Machine Learning Methods to Analyze Children's Vocabulary Learning" (2022). *USF Tampa Graduate Theses and Dissertations.*
https://digitalcommons.usf.edu/etd/10353

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

Effective Statistical and Machine Learning Methods to Analyze Children's Vocabulary Learning

by

Houston T. Sanders

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Mathematics and Statistics
College of Arts and Sciences
University of South Florida

Major Professor: Kandethody Ramachandran, Ph.D.
Chris Tsokos, Ph.D.
Lu Lu, Ph.D.
Howard Goldstein, Ph.D.

Date of Approval:
March 7, 2022

Keywords: Multivariate Adaptive Regression Splines (MARS), ensemble methods, shrinkage methods, tree-based modeling, lexical characteristics

Copyright © 2022, Houston T. Sanders

DEDICATION

I would like to thank my wife, parents, family, friends, and mentors who helped to support and guide my journey. This dissertation would not have been possible without you.

ACKNOWLEDGEMENTS

Throughout my time at University of South Florida I have received a great deal of support and assistance.

I would first like to thank my wife for her support while I undertake this endeavor. She has been invaluable in helping me frame my ideas, reading first drafts, giving me feedback and edits, and help during every stage of my dissertation. Without her, I do not think it would have been possible to complete this work while raising of our two boys.

I would like to thank my advisor, Dr. Ramachandran, for everything he has done to get me here. His focus on precision and mathematical rigor have been pivotal to my understanding of new concepts as I branched out on my own. He always approaches every problem with attention to detail and pushes me to go deeper into the subject. Whenever I get sidetracked, he always helps me consider why the topic will be useful and what my contributions could be. He has taught me that anything we create should be more precise, faster, or give new insights.

I am thankful for my committee members for all of their teaching and support. Dr. Tsokos has always been supportive and does a lot to help student veterans. I learned a lot about academia from him as we worked together on a solutions manual for one of his textbooks. He has a wealth of knowledge to share and always has interesting stories of his work in statistics. Dr. Lu Lu has had a major impact on my understanding of core statistical concepts. She always balanced theoretical statistical concepts with experimental applications to really solidify my understanding. When studying more advanced topics, I always found myself coming back to the

material she taught me, strengthening my understanding. Dr. Goldstein has been very supportive of my work, and I always felt like he was invested in my success. I met him as my wife's mentor and instantly felt like he was my mentor as well. He always checked on my progress and gave feedback, as well as including me in some interesting discussions about research for children's word learning.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Poor methodological and statistical practices can lead to unreliable results. The collaboration between statisticians and researchers can remedy this. Early education intervention research rarely uses advanced statistical techniques. Within early education, vocabulary instruction has been well-studied, yet outcomes continue to be underwhelming. The specialized knowledge and expertise statisticians possess has the potential to enhance word learning research by applying sophisticated analyses not commonly used.

Choosing vocabulary words for instruction can be a daunting task and is highly subjective. In an effort to aid in the selection process, researchers use a word selection framework that groups words into three tiers. Even with words organized into these tiers, there is still considerable variability when selecting words for instruction. There could be other factors related to word learning, and these, combined with a word's tier, would better organize words for instruction. Recent research has been done to examine the lexical characteristics that influence children's word learning and recognition. Multivariate linear regression and stepwise regression are two common statistical analyses used to model these relations. These models can be appropriate in certain situations, but the assumptions they rely on may not be satisfied in the context of word learning models. Interdisciplinary collaboration between statisticians and word learning researchers could lead to more appropriate modeling approaches that better-describe the influence of lexical characteristics on word learning.

The purpose of this three-part dissertation is to advance word learning research by implementing sophisticated statistical techniques that are not commonly used. (i) First, we

introduced and compare the theoretical framework of statistical and machine learning techniques that would be applied to word learning data such as shrinkage methods and ensemble learning. (ii) The performance of these advanced techniques are compared using fit measures and an example subset of the data. We demonstrated why multivariate adaptive regression splines (MARS) is a better choice for a robust word learning model by comparing it to advanced statistical and machine learning techniques, as well as typically used methods by education researchers, such as multivariate linear regression and stepwise regression. (iii) Three word learning datasets were modeled using MARS to examine the relations among lexical characteristics and children's word learning. This was done to see if results were consistent with the first analysis and to determine the differential effects lexical characteristics had on word learning across grade levels.

Words were characterized by various lexical factors including age of acquisition, word frequency, level of concreteness, neighborhood density and phonotactic probability. Compared to multivariate linear regression and stepwise regression results, the different statistical and machine learning techniques performed well, but MARS proved to be superior for its balance of accuracy and interpretability. Results indicated age of acquisition and level of concreteness were the most relevant predictors of word learning. Children had difficulty learning words that were rated older than their age and that were highly abstract. The points at which learning declined appeared to shift as children aged. Examining hinge data, we can determine the threshold for learning words based on this information. Using final models for each grade level, we can predict the number of students expected to learn a given word based on the lexical characteristics. This information can be used to systematically organize vocabulary targets into an optimal sequence for instruction.

CHAPTER ONE:

INTRODUCTION

Statistical errors are common in many fields of research, which has led to calls for action to remedy poor methodological and statistical practices (Sainani et al, 2021; Veldkamp et al, 2014). Sainani and colleagues (2021) make the case for an increase in collaboration between researchers and statisticians, as well as an increase in statistical training for researchers. Statisticians have a variety of tools and skills at their disposal that would be a boon to all manner of research but requires outreach and collaboration. Even when statisticians consult with other departments, it is often beset by challenges due to a lack of direction (Khamis & Mann, 1994). Interdisciplinary research has become more common and statistical rigor has increased for many disciplines, but questionable research practices are still prevalent (John et al, 2012). Fields such as psychology and medicine have made great strides in becoming more statistically rigorous but other research areas are lagging behind (Open Science Collaboration, 2015).

Early education intervention research rarely employs advanced statistical techniques (Snyder et al., 2002). Interventions focused on vocabulary instruction often use simpler analytic methods, such as linear or stepwise regression, which can be inappropriate given the correlated nature of the data. When studying the impact educational interventions have on children's learning, results are typically "messy." Individual differences in children's language and literacy abilities, variability in classroom environments, and the limited ability for experimental control can all lead to data that is heavily skewed. Furthermore, when examining special populations (i.e., children with language impairments), results cannot always be generalized to the greater

population. Understanding data is important for determining the models most appropriate based on certain factors (e.g., linearity and normality, multicollinearity, or homoscedasticity). Advanced statistical methods that do not rely on such assumptions may reveal important information that would be missed otherwise. This means researchers need to know when and how to use these techniques, and understand the underlying assumptions, advantages, disadvantages, and a priori knowledge associated with each technique.

The purpose of this dissertation is to consider statistical learning and machine learning techniques to determine the appropriate method to identify the relevant impact lexical characteristics have on word learning. First a review of relevant literature examining the status quo of word learning will be presented. Second, data from a longitudinal study examining the effects of a supplemental vocabulary intervention on children's word learning will be explored and described. Third, the theoretical frameworks for advanced statistical techniques are introduced and compared to methods commonly used in word learning research. Next, we will compare various advanced statistical techniques used to analyze the influence of lexical characteristics on children's word learning. Finally, we will identify the lexical characteristics predictive of word learning using data from studies investigating the effects of two different vocabulary interventions.

Vocabulary knowledge is crucial for reading comprehension. However, there are considerable differences in children's vocabulary size; those from families with lower socioeconomic status tend to have smaller vocabularies compared to their peers from families with middle and high socioeconomic statuses. These differences are evident as early as four-years-old (Hart & Risley, 1995; 2003). Children with smaller vocabularies are at greater risk for

developing reading disabilities upon entering school. Vocabulary instruction has the potential to reduce the prevalence of reading disabilities.

Vocabulary instruction varies greatly, especially in early education classrooms (Greenwood et al., 2013) and is almost non-existent in classrooms serving children from low-income families (Wright, 2012). Additionally, word selection is a challenging task and relies heavily on teachers to determine the words used for instruction (Gray & Yang, 2015). Several researchers have created word selection frameworks (Beck, McKeown, & Kucan, 2013; Biemiller, 2010; Marzaon & Sims, 2013). Beck and colleagues (2013) organized words into three tiers. Tier 1 words are the basic building blocks of language and commonly used in conversation (i.e., mine, sad, run). Tier 2 words are words that have high academic utility; they are not domain-specific and are words children will encounter while reading and thus impact comprehension (i.e., absurd, construct). Tier 3 words are content specific. While they require instruction, they have very little application outside of that subject area (i.e., hypothesis or biome). Biemiller (2010) organized words in a similar fashion; words that are learned without instruction, words worth teaching, and words to be learned later on. He also distinguished words appropriate for primary and upper-elementary grades. Marzano and Simms (2013) organized words into Tiers 2 and 3. While the Tier 3 words are grouped by grade level, the Tier 2 words are not because these words appear in texts across grade levels making it more challenging to assign these words to a specific grade level. Instruction should focus on Tier 2 words.

While these efforts aid in the selection of words to teach, these frameworks lack a systematic method to identify appropriate instructional targets. In a systematic review of word selection in early childhood vocabulary instruction, Hadley and Mendez (2021) found that, of the studies that used Beck and colleagues' tiered system for word selection, only 41% of the words

were categorized as Tier 2 based on their coding criteria. They also found that several words fell within gray areas, fitting into more than one tier. They also found that the application of the tiered system varied greatly. To this point Hadley and Mendez posed an interesting question: how do word tiers vary by age? Would a Tier 2 word for a preschooler also be a Tier 2 word for a child in fifth grade? A word's tier cannot be the only deciding factor used to select targets for instruction.

There may be other factors that influence word learning that could be used to better-organize words for instruction. Researchers have examined the lexical characteristics that may facilitate word learning, but the methods used to model these relations may under- or overestimate their contributions to children's vocabulary acquisition. Increased collaboration between statisticians and educational researchers is needed. By interacting with researchers, statisticians can provide the tools and understanding needed to reveal important relations between lexical characteristics and word leaning.

By demonstrating the value of using advanced methods to analyze word learning, researchers will be better equipped to interpret results that have the potential to impact vocabulary instruction. Outcomes illustrate the importance of promoting collaboration between statisticians and vocabulary intervention researchers by increasing the statistical rigor in this area. This will lead to an enhanced understanding of word learning outcomes which can be used to develop a systematic method for selecting instructional vocabulary targets. This would lead to the development of an algorithm used for word selection. This algorithm will be based on relevant predictors and allow for a flexible approach to selecting words that are developmentally appropriate for children. Better-organized vocabulary targets would lead to an increase in

learning, potentially closing the vocabulary gap and reducing the prevalence of literacy disabilities among vulnerable populations.

CHAPTER TWO:

LITERATURE REVIEW

Advanced Techniques in Educational Research

While the majority of educational intervention researchers make limited use of advanced statistical techniques, two recent meta-analyses highlighted the few studies that used advanced analytic techniques. Kormaz and Correia (2019) performed a review of machine learning in education research with a focus on methods such as support vector machines, Bayesian networks, fuzzy logic, and decision trees. They found that the use of machine learning techniques was trending upwards but were still relatively small in comparison. Nájera and Mora (2017) reviewed education applications of data mining and machine learning. Because educational research is still very reliant on simpler analysis methods, such as regression, their goal was to show how machine learning methods can solve difficult and interesting problems. They found examples using decision trees, neural networks, naïve Bayes, k-nearest neighbors, logistic regression, and support vector machines and gave a brief overview of how to choose a machine learning model.

Beyond basic linear regressions, mixed effects models and multilevel modeling have been used to test the nested interaction between parameters. Kelley and colleagues (2020) used a $2 \times 2 \times 9$ multilevel model to examine the extent to which a vocabulary program impacted preschoolers' sophisticated word learning. Results revealed significant interactions between condition (treatment vs. control), time (pre-intervention vs. post-intervention), and instructional book.

Lindl and colleagues (2020) introduced mixed effects modeling and multilevel models as well as structural equation models (SEM) and random forests. Structural equation models look for latent connections that are represented with a path model. Each of the models were introduced with a simple dataset and demonstrated using an example. Harlaar and colleagues (2007) used Cholesky decomposition and compared it to the variance-covariance matrices to determine the factors among twins that impacted word learning. They suggested that SEM would serve as an alternative model that would perform equally well.

Some researchers have adopted random forests for modeling educational data. One such study was to use survey data completed by 30 English as a second language instructors who rated the complexity of 7,000 words. The survey data was combined with word frequencies from a database of 50,000 words and was modeled with a random forest to rate their complexity (Sohsah et al, 2015). One issue they ran into was the unbalanced nature of word usage, which made modeling accurately difficult. Random forests have also been used to model student achievement using demographic survey data to predict outcomes based on the Programme for International Student Assessment (PISA; Güre et al, 2020).

The prior two studies also used neural networks to analyze their data. Sohsah and colleagues (2015) used a 2-layer feed forward neural network and Güre and colleagues (2020) used an artificial neural network (ANN), as well as a multilayer artificial neural network (MLANN). Neural networks have been used to predict student's GPA, academic retention, and degree completion based on a variety of measures such as: Attention Network Test (ANT; Fan et al, 2002), Learning Strategies Questionnaire (LASSI; Weinstein & Palmer, 2002), and many others (Musso et al, 2020).

Support vector machines (SVM) is another technique that has been used in educational research. Again, Sohsah and colleagues (2015) modeled their word frequency data with support vector machines to classify the complexity of English words. SVM was used by researchers to take 49 context factors from student surveys for 2,646 high achieving students and 1,369 low achieving students and classify students based on the 2015 PISA reading literacy test (Dong & Hu, 2019). The data from students participating in the Progress in International Reading Literacy Study (PIRLS) from 2016 was modeled using SVM to classify and predict high and low proficiency readers (Chen et al, 2020).

The prior study also used logistic regression to classify the students based on PIRLS. Another study took 986 words and performed dimensional reduction using principal component analysis on the words and then performed logistic regression using information about child participants (i.e., sex, age, quantifiers of vocabulary) and a created a numeric representation that described specific words children knew to predict whether they would learn a given word (Beckage et al, 2015). Principal component analysis for dimensional reduction has been used by other researchers, including Yap and colleagues (2012) who reduced the dimension of their data on visual word recognition before modeling it with multivariate linear regression.

Gradient boosting is rarely used in educational research. During this literature review, two studies were found to employ this technique. Fifty-one lexical features among eight groupings were used to model the complexity of words using gradient boosting trees (Agarwal & Chatterjee, 2021). The PIRLS 2016 research by Chen and colleagues (2020) used extreme gradient boosting as one of the methods to classify and predict high and low proficiency readers.

They also modelled their data using decision trees. Another study used decision trees to predict student's performance from the PIRLS data to model the influence on reading ability (Alivernini, 2013).

Classification and regression trees (CART) were used with data from PISA, Trends in Mathematical and Science Study (TIMSS), and PIRLS to compare student achievements from different countries and determine where countries could improve student education (Depren, 2018). This study also used multivariate adaptive regression splines (MARS) to model datasets and compared them to CART. They found that MARS outperformed CART.

From the literature it can be seen that some advanced techniques like gradient boosting, random forests, and MARS are being implemented in educational research. Though it exists, the scale is limited and is mostly used to classify student achievement based on large scale international testing. Specifically, those techniques are not being used in early childhood research. In a review of early intervention studies conducted over a ten-year time span, the majority of studies used univariate parametric analyses (Snyder et al., 2002). Only 23 studies out of 450 included in this review used advanced techniques. Additionally, these advanced techniques are not used to analyze vocabulary acquisition. Many researchers examining the impact lexical characteristics have on children's word learning use both multivariate linear regression and stepwise regression (Gray, 2004; Morrison & Ellis, 2000; Stoel-Gammon, 2010; Storkel, 2009). While these methods are well-known and commonly used, results may not be reliable and can over- or under-estimate the effects of lexical characteristics on word learning. Advanced statistical techniques could better-describe word learning data.

Importance of Vocabulary Instruction

Oral language, including children's vocabulary knowledge, is an essential prerequisite for reading comprehension (Anderson & Nagy, 1991; Elleman et al., 2009; Snow, Burns, & Griffin, 1998; Taffe et al., 2009). Early childhood is a critical time in children's oral language development, but language development varies greatly among children. Those from families with a low socioeconomic status tend to have smaller vocabularies compared to their peers from families with middle and high socioeconomic statuses, and these differences are evident as early as four-years-old (Hart & Risley, 1995; 2003). Children with limited oral language skills are at greater risk for developing later reading disabilities (Catts, Fey, Zhang, & Tomblin, 1999; Scarborough, 1998; Sénéchal, Oullette, & Rodney, 2006). Vocabulary instruction is key to closing this achievement gap.

Despite the well-established role of vocabulary instruction in children's development of oral language and reading skills, little is known about what words to teach and when. Vocabulary instruction has been well-studied (Beck & McKeown, 2007; Coyne et al., 2007; Goldstein et al., 2017; Kelley et al., 2020; Justice et al., 2005; McKeown & Beck, 2014; Storkel et al., 2017). Researchers have identified instructional practices that promote word learning including explicit instruction, using child-friendly definitions, providing multiple contexts for the words, and connecting vocabulary words to real-world examples (Beck, McKeown, & Kucan, 2013). Even though vocabulary acquisition is of great interest, results continue to be underwhelming (Wasik et al., 2016). Additionally, there is great variability in vocabulary instruction and word selection, especially in early childhood classrooms (Greenwood et al., 2013). Several researchers have attempted to organize vocabulary words for instruction (Beck, McKeown & Kucan, 2013; Biemiller, 2010; Marzano & Pickering, 2005), but these systems for grouping words lack a

standardized, systematic approach to word selection. There may be other factors that influence word learning outside of instructional methods that could be used to organize words into a developmentally appropriate sequence for instruction. Individual characteristics like part of speech, imageability, or frequency of use may contribute to the learnability of a word.

Lexical Characteristics

Word frequency measures a word's frequency of use in a given language, in this case American English. Words with a high frequency are used more often than words with a lower frequency. There are several measures for word frequency for American English (i.e., Francis & Kučera, 1982). The $SUBTLEX_{US}$ word frequency measure is a corpus that can be accessed online and provides frequencies for spoken language that approximates everyday language use (Brysbaert & New, 2009). The values represent the frequency per million words. The corpus contains 51 million words and is based on American English subtitles from movies and television shows.

Age of acquisition is the age at which a person learns a particular word. Kuperman and colleagues (2012) compiled age of acquisition (AoA) ratings for 30,000 words selected from the $SUBTLEX_{US}$ corpus. The ratings were obtained by asking 1,960 individuals to rate the age at which they learned each word, defined by understanding the word feature when used by others but not necessarily used by themselves. While this may seem like a difficult task, researchers have found that adult ratings of age of acquisition are accurate (Gilhooly & Gilhooly, 1980; Gilhooly & Logie, 1980).

Level of concreteness is defined by Brysbaert and colleagues (2014) as imageable or abstract. Imageable words are things that can experienced through the five senses (e.g., rock, jump). Abstract words cannot be experienced, and their meanings must be defined by other

words (i.e., freedom, justice). Ponari and colleagues (2018) examined how children learn abstract words and found that young children rely heavily on emotional valence to learn abstract words until approximately ages eight or nine, then shifts to relying more on linguistic information. As children age, their capacity to learn more abstract words increases, utilizing earlier acquired words that may be more concrete as a foundation to build upon.

Phonological neighborhood density describes the organization of phonetically similar words in the mental lexicon. The neighborhood for a word is made up of a group of words that differ by one sound substitution, deletion, or addition. For example, the word "aid" has a neighborhood density of 21634.85 meaning that it has over 20,000 phonetically similar neighbors (i.e., aim, paid, maid) whereas the word "appearance" had a neighborhood density of 0 meaning that it does not have any other phonetically similar neighbors. According to Luce and Pisoni's (1998) Neighborhood Activation Model, the frequency with which words are used, and the density of the neighborhood, effect spoken word recognition, discrimination, and the amount of time needed to find and produce a word.

Phonotactic probability is the frequency of phonological segments and sequences of phonological segments that occur in words in a given language (Vitevitch & Luce, 2004). To accomplish this, the sum of the log token frequency of words containing position-specific phonemes in that segment is divided by the log token frequency of all words containing the segment. Common sound sequences have a higher phonotactic probability than those with combinations not as common. Phonotactic probability and neighborhood density are significantly correlated with one another (Vitevitch et al, 1999).

Prior research has examined the effect these lexical characteristics have on word learning in both children and adults (Newman & German, 2002; Hadley et al., 2021; Hoover et al., 2010;

McDonough et al., 2011; Storkel, 2001; Storkel et al., 2006). Newman and German (2002) found words with typical stress patterns (i.e., 13ormal1313), high in frequency, and low in neighborhood density and age of acquisition (words learned at a younger age) were easier for children to name. Hoover and colleagues (2010) found preschoolers learned words with common sound sequences in dense neighborhoods, and words that contained infrequent sound sequences in sparse neighborhoods. McDonough and colleagues (2011) found a relation between age of acquisition and word imageability (or concreteness). Words that were more concrete were learned earlier. Similarly, Hadley and colleagues (2021) found a word's imageability significantly predicted preschool children's word learning. Results from their mixed-effects models found that imageability explained 34% of variance across words. Storkel (2001) examined the effect phonotactic probability had on preschoolers' word learning and found that preschool children acquired words with common sound sequences faster than words with rare sound sequences. Furthermore, in a study of adult word learning, Storkel and colleagues (2006) found phonotactic probability contributed to word learning in adults and neighborhood density promoted the integration of new and existing lexical representations. It is important to note that the researchers in these studies controlled for phonotactic probability and neighborhood density by creating pseudo-words, and because of this, the results may over generalize the effects these characteristics have on word learning compared to alternative word learning studies where these are not controlled (Hoover et al, 2021; Storkel, 2001; Storkel et al., 2006).

Identifying the lexical characteristics most relevant to children's word learning can facilitate the selection of vocabulary targets used for instruction. Using these relevant factors, we can create a systematic approach to word selection. By utilizing these lexical characteristics for word organization, it is possible to create a more unified, developmentally appropriate sequence

of vocabulary targets used for instruction. This will facilitate the selection process when planning vocabulary instruction. Enhancing vocabulary instruction has the potential to reduce the prevalence of reading disabilities especially among children from vulnerable populations closing the achievement gap.

CHAPTER THREE:

DATA DESCRIPTION AND EXPLORATORY ANALYSIS

Description of Data

Participants

Word learning outcomes were collected for approximately 350 students from first, second, and third grade classrooms who took part in a longitudinal study investigating the effects of a supplemental intervention that taught academic vocabulary words. Students attended two elementary schools that served primarily low-income families. Over 90% of students qualified for free or reduced lunch.

Supplemental Vocabulary Intervention

The Independent Lexical Instruction and Development (ILIAD) supplementary Tier 2 vocabulary program (Goldstein et al., 2017) was a longitudinal study spanning three years. The intervention occurred four days a week and was set up as a listening center where students would follow along with a pre-recorded lesson that included a read aloud from books used in the core curriculum, the Open Court series. Lessons were scripted and included opportunities to interact and respond and provided multiple opportunities for students to interact with the target words.

Across each grade level a total of six Tier 2 words: two nouns, two verbs, and two adjectives were taught each week. In first and second grades an additional Tier 1 anchor word from the Open Court series was included to connect the lesson to the classroom curriculum. Tier 2-word selection for the intervention followed the criteria set forth by Beck and colleagues

(2002). The Academic Word List (Coxhead, 2000) included a list of words derived from a variety of college-level texts could be categorized as Tier 2 words. A total of 377 words were included. Each of the words chosen for instruction had to be illustrated, defined, and fit into existing stories. Because of these constraints researchers were running out of words for third grade. The Living Word List (Dale & O'Rourke, 1976) was used to supplement word selection.

Learning Outcome Measures

The learning outcomes were derived from two subtests of a researcher-made measure that was administered every 4-5 weeks and included an expressive labeling probe that required students to identify the target word when presented with a trained picture stimulus along with or without the definition. The decontextualized definition probe required students to provide the meaning of the target word without additional contextual support (Goldstein et al., 2017). Students were prompted with *"Tell me everything you know about ___."* If they responded with one attribute of the word, they were prompted a second time with *"Tell me something else about___."* A correct response included a definition, a synonym or brief description of the word. Originally, this measure was scored on a three-point scale, 0 for not learned, 1 for partial knowledge, and 2 for full knowledge. Because we were interested in the number of children who learned each word, we collapsed partial and full knowledge into a binary scale. The revised scale for this secondary analysis was 0 for not learned and 1 for learned.

Coding of Lexical Characteristics of Words

A total of 377 target vocabulary words were characterized for analysis based on available database estimates of their individual word frequency, age of acquisition, phonological phonotactic probability, neighborhood density, and level of concreteness (see Table 3.2 for mean, standard deviations, and ranges for each lexical characteristic). Word frequency values,

phonotactic probability, and phonological neighborhood density counts were obtained from the Irvine Phonotactic Online Dictionary version 2.0 (Vaden, Halpin & Hickok, 2009), which reports frequency measures from the SUBTLEXus corpus. Concreteness level ratings were derived from a database of 37,058 English words developed by Brysbaert et al. (2014). Age of acquisition ratings for 30,121 English content words were reported by Kuperman et al. (2012). In some instances, the targeted vocabulary word was a derivation and not included in the databases. When this occurred the values for the base or root word were used instead.

Each database was either available for download as or was converted to Excel files to streamline data collection. Using the search and retrieval functions in Excel, the various databases were searched for all 377 target words. A secondary matching function and random searches by the researcher were done to ensure correct words and values were reported from each database.

Data Analysis

When determining which statistical methods are appropriate for analysis, it is important to consider the data. Exploratory data analysis was performed for the total dataset and for each of the grade levels. Descriptive statistics for word learning outcomes based on decontextualized learning and expressive labeling tasks can be found in Table 3.1. The percentages in the table represent the percentage of students that learned each vocabulary word. This preliminary look at the data shows that the average word learning ranges from 22% to 38% for both decontextualized learning and expressive labeling but the medians are below the mean for every case. For both learning outcomes, grade 3 has a lower maximum learning and standard deviation for the words than other grade levels.

Table 3.1. *Descriptive Statistics for Word Learning*

	Expressive					Decontextualized				
	M	SD	m	Min	Max	M	SD	m	Min	Max
Combined	29%	23%	23%	0%	99%	29%	26%	19%	1%	99%
1st grade	22%	24%	13%	0%	98%	26%	29%	14%	1%	99%
2nd grade	38%	24%	32%	7%	99%	38%	28%	27%	3%	97%
3rd grade	27%	13%	26%	3%	61%	22%	15%	18%	3%	74%

Note. M= mean, SD= standard deviation, m= median, min= minimum value, max= maximum value.

Descriptive statistics for model variables can be found for decontextualized learning and expressing tasks for the full dataset, first grade, second grade, and third grade in Table 3.2. Neighborhood density and word frequency are highly right skewed for every grade level, for both decontextualized and expressive learning. Boxplots for these lexical characteristics are in Figure 3.1. Because the scales for the variables are very different, each variable was standardized for comparison.

Before any inferences were considered, exploratory data analysis was completed for the ILIAD dataset for the full dataset and each respective grade. Figure 3.2 shows the word learning in descending order for each subset of the data for decontextualized learning and expressive tasks, that is, the y-axis represents the percentage of students that learned a given word and the x-axis is the i^{th} word ordered by the percentage of students that learned the word. By ordering the words based on learning, the graphs illustrate the level the learning that is occurring and we can observe that the upper threshold for learning is much lower for the third graders based on both learning outcome measures. The full dataset, first grade, and second grade have similar learning trends with each other. As different methods model the data, this will be important to keep in mind as the influence of lexical characteristics are being tested as predictors of the word learning outcomes.

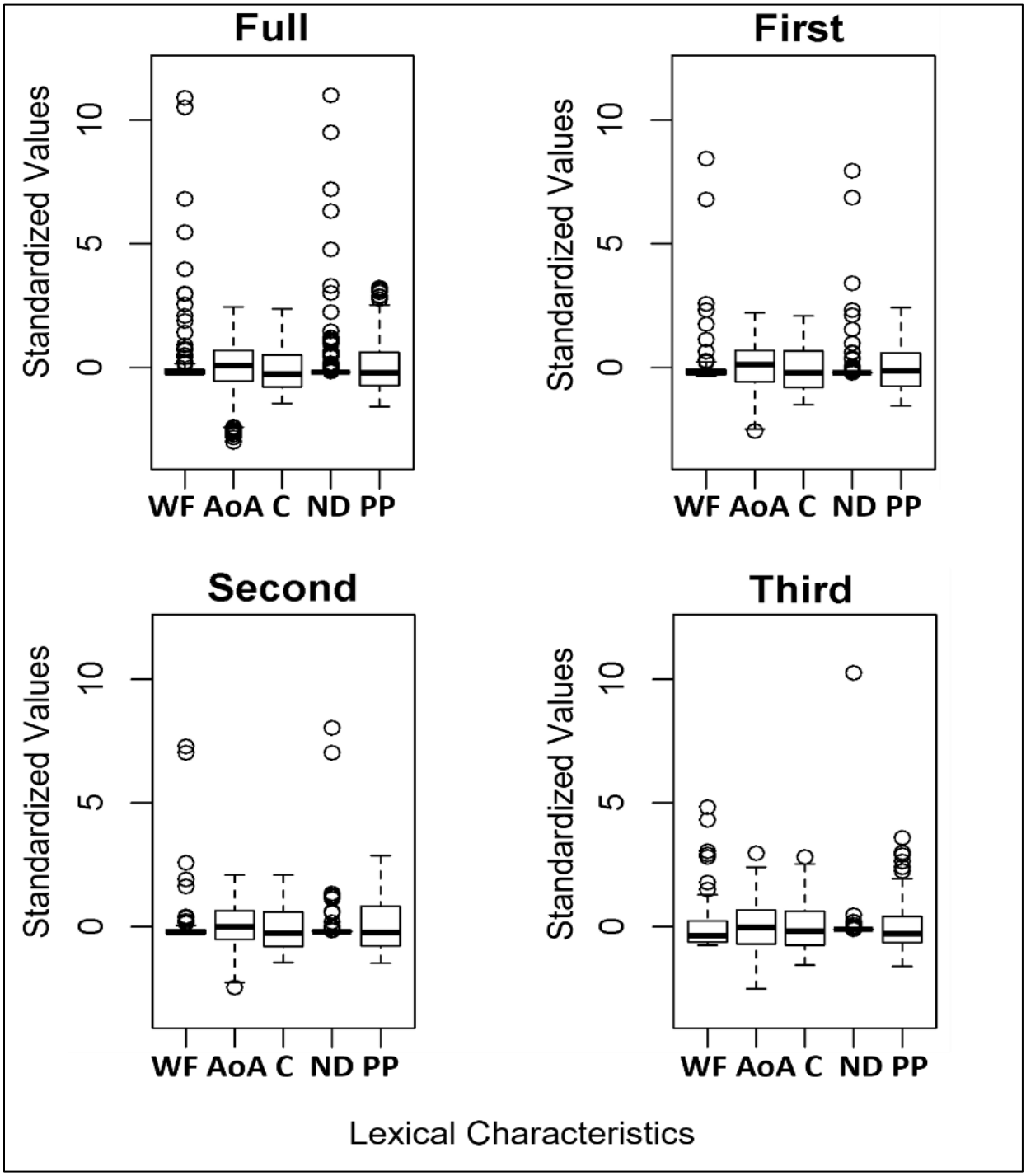
Table 3.2. *Descriptive Statistics for Model Variables (Lexical Characteristics)*

Full Dataset	M	SD	m	Min	Max	Skew
AoA	9.14	2.18	9.33	2.6	14.5	-.58
N_Den	1115.78	6184.45	4.32	0	69210.62	8.08
Conc_Mean	2.78	.93	2.54	1.43	5	.87
Phon_Prob	.23	.13	.21	.02	.66	.85
SUBTLwf	19.29	71.82	5.16	.02	801.82	8.26
First Grade (n= 143)						
AoA	8.80	2.17	9.06	3.25	13.61	-.46
N_Den	1845.32	8474.12	6.49	0	69210.62	6.23
Conc_Mean	2.96	.97	2.76	1.50	5	.64
Phon_Prob	.22	.12	.21	.03	.52	.57
SUBTLwf	19.29	57.94	6.90	.27	509.37	6.67
Second Grade (n= 126)						
AoA	8.63	2.29	8.63	3	13.41	-.23
N_Den	1106.11	5560.46	8.53	0	45721.92	6.92
Conc_Mean	2.89	1.00	2.63	1.46	4.97	.72
Phon_Prob	.24	.15	.21	.02	.66	.76
SUBTLwf	31.37	105.91	7.63	.02	801.82	6.07
Third Grade (n= 108)						
AoA	10.30	1.41	10.25	6.75	14.5	.15
N_Den	771.50	6666.04	1.41	0	69210.62	10.02
Conc_Mean	2.39	.62	2.29	1.43	4.15	.81
Phon_Prob	.24	.14	.21	.03	.72	1.37
SUBTLwf	4.82	6.39	2.46	.08	35.65	2.57

Note: M= mean, SD= standard deviation, m= median, min= minimum value, max= maximum value, n= number of words, AoA= age of acquisition, N_Den= neighborhood density, Conc_Mean= level of concreteness, Phon_Prob=phonotactic probability, SUBTLwf= word frequency.

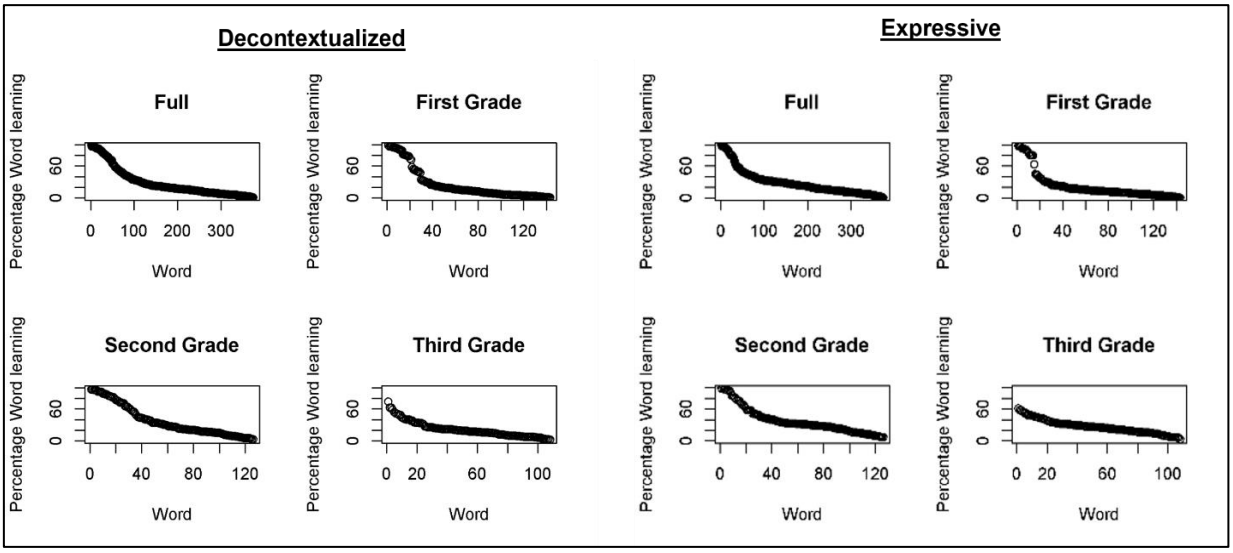
One of the central assumptions of most models is homoscedasticity. This assumes that the error terms, or noise, is the same across the same independent variables. Homoscedasticity was tested for the datasets using the Breusch and Pagan technique (1979) that uses Lagrangian multipliers to find a test statistic. This was calculated in R using the lmtest package (Zeileis & Hothorn, 2002). The full dataset and first grade dataset were found to be heteroscedastic with Breusch-Pagan test scores of 28.16 and 24.96, respectively. The second grade and third grade datasets were found to be homoscedastic with scores of 6.20 and 10.85. This may impact some models that rely on homoscedasticity as an assumption.

Figure 3.1. *Box Plots for Model Variables (Lexical Characteristics)*



Note. WF= Word Frequency, AoA= Age of Acquisition, C= Concreteness, ND= Neighborhood Density, PP= Phonotactic Probability.

Figure 3.2. *Decontextualized and Expressive Word Learning by Grade (Descending)*



Note. Word learning in descending order for each subset. The y-axis represents the percentage of students that learned a given word and the x-axis is the ith word ordered by the percentage of students that learned the word.

Multicollinearity

Multicollinearity was checked for the data to determine if it would impact model creation. There are several methods to check collinearity such as checking the signs of coefficients, comparing coefficients to prior knowledge, test deletion of data to check impact on models, check correlations between all predictor variables, or calculating variance inflation factors (Draper, N. R., & Smith, H., 1998). The simplest approach is generally to produce the correlation matrix R between the predictor variables (Tamhane, A., & Dunlop, D., 2000).

The Pearson correlation coefficient, or Pearson's r, was used to find the linear correlations between the lexical characteristics. The correlation matrix is defined as

$$R_r = (r_{ij}) = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix},$$

where $r_{ij} = \sigma_{ij}/\sigma_i\sigma_j$ (Rencher & Schaalje, 2008). If

$$D_\sigma = [diag(\Sigma)]^{1/2} = diag(\sigma_1, \sigma_2, \dots, \sigma_p)$$

then P_ρ can be found using

$$R_r = D_\sigma^{-1}\Sigma D_\sigma^{-1},$$

where

$$\Sigma = cov(y) = \begin{pmatrix} 1 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & 1 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & 1 \end{pmatrix}.$$

Table 3.3 shows the correlations between independent variables for the full data, first grade, second grade, and third grade, respectively. The correlations were calculated in Rstudio using the base R language (R Core Team, 2019) For the full data, there exists some correlation between all the variables, which stronger negative correlations between age of acquisition and word frequency ($r = -.4$) and between age of acquisition and level of concreteness ($r = -.56$). For the first-grade data, age of acquisition has a strong negative correlation with level of concreteness ($r = -.62$) and a moderate negative correlation with word frequency ($r = -.36$). Age of acquisition has a strong negative correlation with word frequency ($r = -.44$) and level of concreteness ($r = -.51$), word frequency has a strong correlation with neighborhood density ($r = .53$), and moderate correlation for most other variable combinations. The third-grade data is mostly uncorrelated other than age of acquisition and word frequency which are negatively correlated ($r = -.51$). The correlation among the variables was expected because prior studies have shown a link between lexical characteristics (Hoover et al, 2010; Vitevitch et al, 2004; Storkel, 2004). These strong correlations for the lexical characteristics may be important while

selecting the best model, but Pearson correlations are not reliable with skewed data (Zou et al, 2003), which ILIAD is, so other measurements were considered. Figure 3.3 displays graphical representation of the correlations among the variables for each subset of the data.

Table 3.3. *Correlation between Variables*

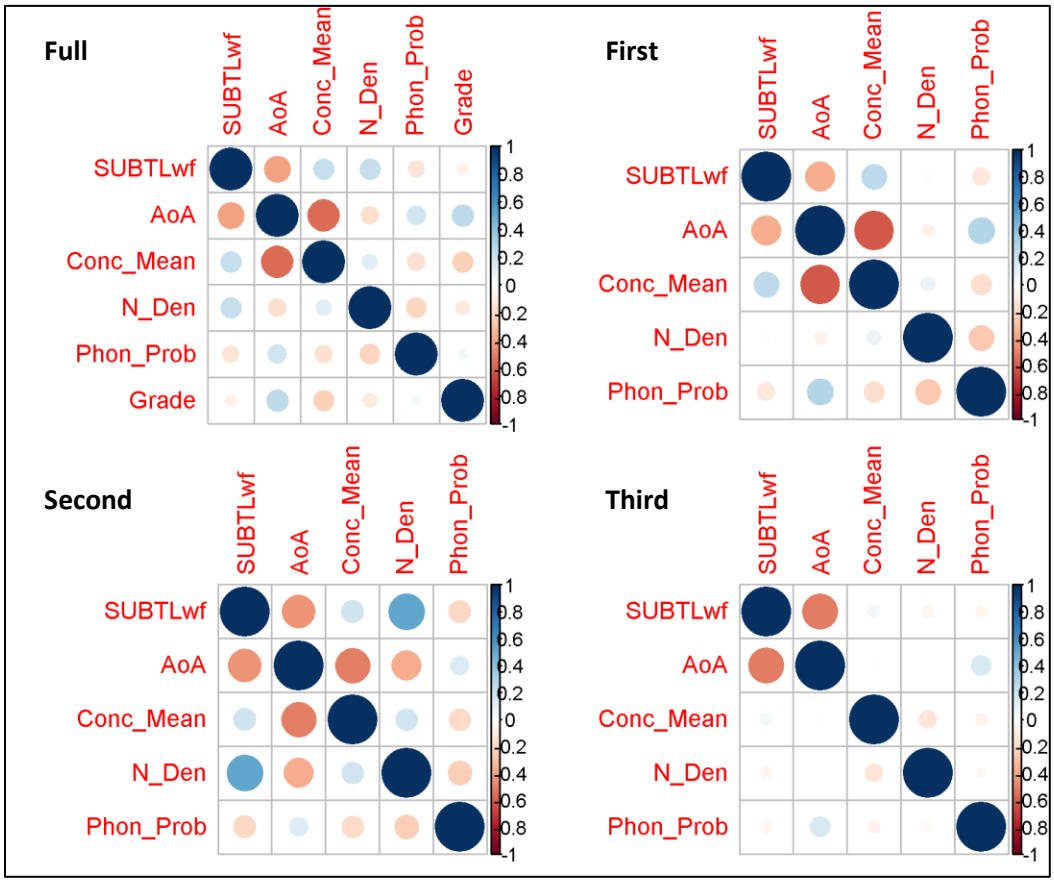
Full Dataset	W Freq	AoA	Concrete	Nden	Phon Prob
W Freq	1				
AoA	-.4	1			
Concrete	.24	-.56	1		
Nden	.24	-.18	.13	1	
Phon Prob	-.15	.20	-.17	-.22	1
First Grade					
SUBTLwf	1				
AoA	-.36	1			
Concrete	.27	-.62	1		
Nden	-.02	-.06	.08	1	
Phon Prob	-.13	.29	-.17	-.26	1
Second Grade					
SUBTLwf	1				
AoA	-.44	1			
Concrete	.21	-.51	1		
Nden	.53	-.36	.20	1	
Phon Prob	-.20	.14	-.19	-.24	1
Third Grade					
SUBTLwf	1				
AoA	-.51	1			
Concrete	.06	-.01	1		
Nden	-.06	.01	-.13	1	
Phon Prob	-.05	.16	-.06	-.04	1

Note. SUBTLwf= word frequency, AoA= age of acquisition, Concrete= level of concreteness, N_Den= neighborhood density, Phon_Prob= phonotactic probability

Next multicollinearity was tested using variance inflation factors (VIF) using the olsrr package using R (Hebbali, 2020). Multicollinearity is tested by considering the extent of its singularity $X'X$, or how close the determinant is to zero. Here $(n-1)^{-1}X'X = R$ will be the

correlation matrix if the $x_j's$ are suitably standardized. VIF uses the diagonal elements of R^{-1} and as these elements increase, the associated variance increases for $\hat{\beta}_j$.

Figure 3.3. *Correlation between Variables (Lexical Characteristics)*



Note. Graphical representation of correlation between variables. SUBTLwf= word frequency, AoA= age of acquisition, Conc_Mean= level of concreteness, N_Den= neighborhood density, Phon_Prob= phonotactic probability.

$$VIF_j = \frac{1}{1 - r_j^2} = \frac{1}{Tolerance}, \qquad j = 1, 2, \ldots, k,$$

where r_j^2 is the multiple correlation coefficient obtained when the jth predictor variable column X_j is regressed against all other predictors X_i with $i \neq j$. r_j^2 will be close to 1 and VIF will be large if X_j is approximately linearly dependent on the other predictor variables. Based on this,

lower VIF values means that the data has lower multicollinearity (Tamhane & Dunlop, 2000). Values between 1 and 5 are considered small to moderate but VIF is unable to distinguish among several coexisting near dependencies and a lack of a meaningful boundary between levels of multicollinearity (Belsley et all, 2005). The variance inflation factor for each variable by dataset can be found in Table 3.4 and based on the results, there will likely be a moderate impact due to correlation and multicollinearity.

Table 3.4. *Variance Inflation Factors*

Variable	Full	First Grade	Second Grade	Third Grade
W Freq	1.24	1.16	1.55	1.37
AoA	1.70	1.83	1.63	1.39
Concrete	1.48	1.63	1.37	1.03
Nden	1.11	1.08	1.46	1.02
Phon Prob	1.09	1.17	1.10	1.04

Nonlinear Relationships and Normality

After finding the linear relationships between parameters, nonlinear relationships were considered because of the data skewness and for the sake of being thorough. One such method used was using information theory to calculate the mutual information between variables. Mutual information $I(X;Y)$ can be rewritten as

$$I(X;Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} = H(Y) - H(X|Y)$$

where $H(X)$ represents the entropy of X (Cover, 1999). Mutual information was calculated between variables and does not show much of a relationship between variables. The maximal information coefficient (MIC) belongs to a larger class maximal information-based

nonparametric exploration (MINE) and was calculated by finding when $H(X) = H(Y) = H(X,Y)$. MIC is an equitable statistic, meaning it can give similar scores to equally noisy relationships regardless of the type of relationship (Reshef et al, 2014). The MIC values will always fall between 0 and 1 (Reshef et all, 2011) with 0 representing statistically independent variables. These values were computed using the 'minerva" package in the R environment (Albanese et al, 2013).

Table 3.5 shows the MIC values for the mutual information between each variable for the full dataset, first grade, second grade, and third grade, respectively. For the full dataset, age of acquisition shares a moderate amount of information with word frequency (MIC=.39), level of concreteness (MIC=.31), and neighborhood density (MIC=.31). For the first grade data age of acquisition shares a moderate amount of information with word frequency (MIC=.35), level of concreteness (MIC=.40), and neighborhood density (MIC=.35) and word frequency with neighborhood density (MIC=.33). Age of acquisition shares a moderate amount of information with word frequency (MIC=.43), level of concreteness (MIC=.34), and neighborhood density (MIC=.35) for the second-grade data. In the second-grade data there is also moderate mutual information between word frequency and neighborhood density (MIC=.39) and phonotactic probability and neighborhood density (MIC=.30). For the third-grade data, word frequency is moderately related to age of acquisition (MIC=.45), neighborhood density (MIC=.41), and phonotactic probability (MIC=.31). Age of acquisition shares moderate information with neighborhood density (MIC=.30) and phonotactic probability (MIC=.31).

This shows that there is a moderate amount of mutual information between variables for each subset of the ILIAD dataset. Mutual information has been demonstrated to be a strong statistic that can be used for feature selection and can be used for split decisions in regression

trees (Zaffalon & Hutter, 2002; Fleuret, 2004; Hoque et al, 2014). Mutual information is not reliant of any particular model, as much as the assumptions of normality, or needing to reorder or rank the outcomes. Mutual information can take a moderate amount of data to feel confident in the MIC measurements. To be more confident in determined relationships between variables, more tests were done to verify the findings.

Table 3.5. *Maximal Information Coefficients (MIC)*

MIC Full	W Freq	AoA	Concrete	Nden	Phon Prob
W Freq	1				
AoA	.39	1			
Concrete	.18	.31	1		
Nden	.34	.31	.21	1	
Phon Prob	.18	.21	.17	.22	1
First					
W Freq	1				
AoA	.35	1			
Concrete	.25	.40	1		
Nden	.33	.35	.27	1	
Phon Prob	.22	.24	.25	.28	1
Second					
W Freq	1				
AoA	.43	1			
Concrete	.27	.34	1		
Nden	.39	.35	.26	1	
Phon Prob	.26	.28	.23	.30	1
Third					
W Freq	1				
AoA	.45	1			
Concrete	.25	.25	1		
Nden	.41	.30	.23	1	
Phon Prob	.31	.31	.19	.24	1

The next approach for testing the relationship between the variables was to consider a nonparametric ranked correlation measure. To do this, normality of the data was first tested. Using Shapiro-Wilk test for normality and quantile-quantile (Q-Q) plots the data was determined to be nonparametric, specifically neighborhood density and word frequency. To compute the

Shapiro-Wilks test for normality, the denominator D for the test statistic must first be calculated (Conover, 1998)

$$D = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

and order the sample from smallest to largest.

$$X^{(1)} \leq X^{(2)} \leq \cdots \leq X^{(n)}.$$

$\bar{X}$ is the sample mean for the data and $X^{(i)}$ denotes the ith order statistic. The test statistic for this method is

$$T = \frac{1}{D}\left[\sum_{i=1}^{k} a_i\left(X^{(n-i+1)} - X^{(i)}\right)\right]^2.$$

This test statistic is commonly referred to as the W test.

The results for the Shapiro Wilk's tests for each variable based on the associated dataset are in table 3.6. The W scores displayed within the table describe how closely the data follows a normal distribution, with values closer to zero being less normally distributed. The significance for each statistic represents the hypothesis test that the variable follows a normal distribution. Based on this, age of acquisition was found to be normal for the second and third grade data subsets and all others were non-normal. Age of acquisition, level of concreteness, and phonotactic probability were generally close to being normal for each dataset but word frequency and neighborhood density differed considerably. This can be verified visually by Figures 3.4 – 3.8 display the quantile-quantile (QQ) plots for each dataset for age of acquisition, neighborhood

density, level of concreteness, phonotactic probability, and word frequency, respectively. Graphically, the plots agree with what was found using the Shapiro Wilk test.

Table 3.6. *Shapiro-Wilk's W*

	W Freq	AoA	Concrete	Nden	Phon Prob
Full Model	.22*	.97*	.91*	.17*	.94*
First Grade	.28*	.98*	.93*	.23*	.95*
Second Grade	.27*	.98	.90*	.20*	.93*
Third Grade	.69*	.99	.94*	.09*	.88*

Figure 3.4. *Quantile-Quantile Plot for Age of Acquisition*

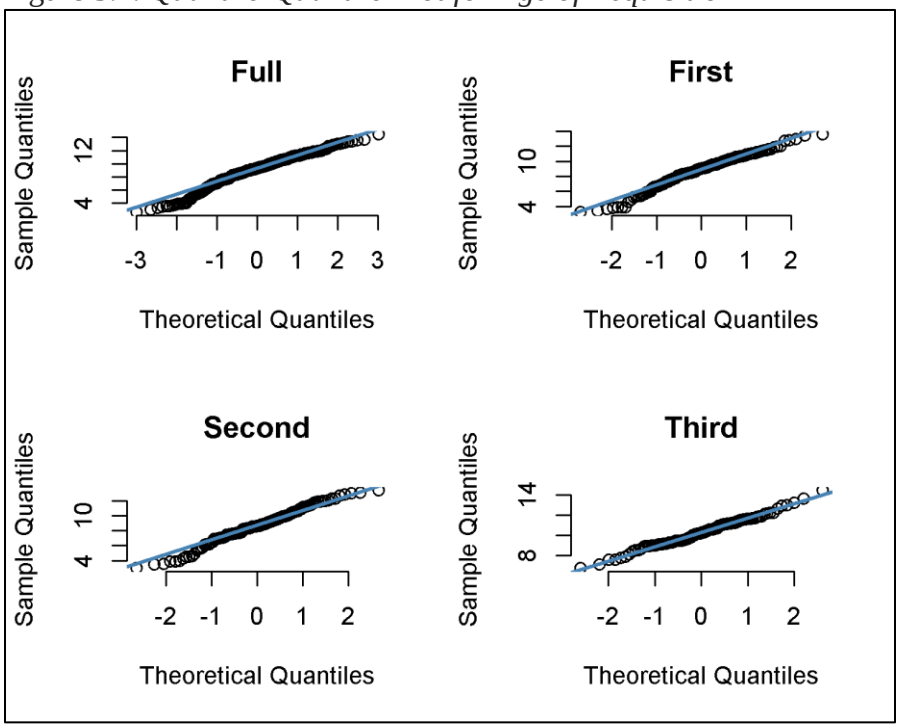


Figure 3.5. *Quantile-Quantile Plot for Neighborhood Density*

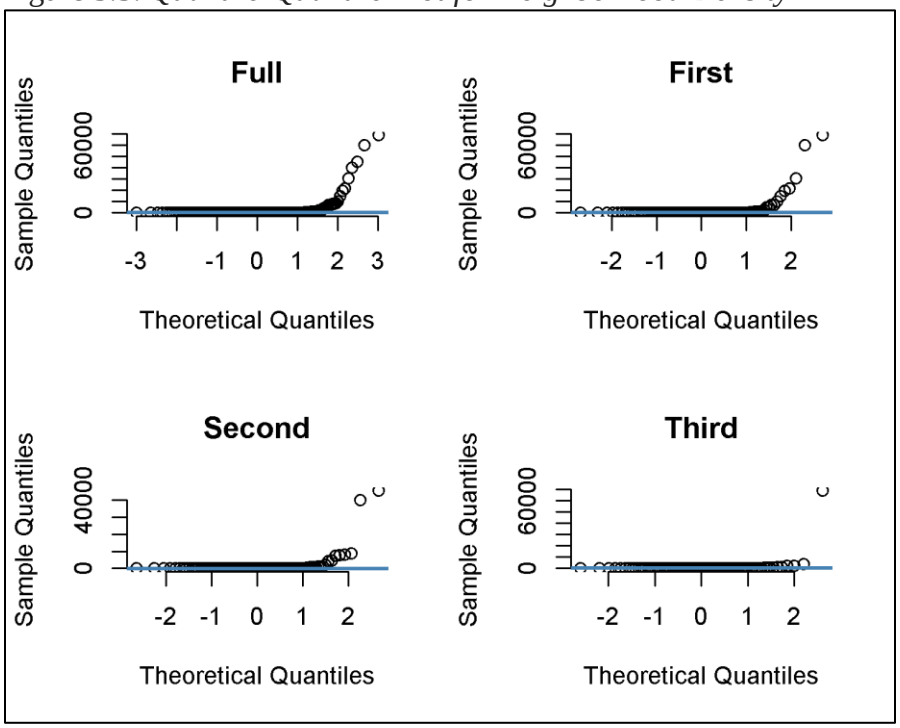


Figure 3.6. *Quantile-Quantile Plot for Concreteness*

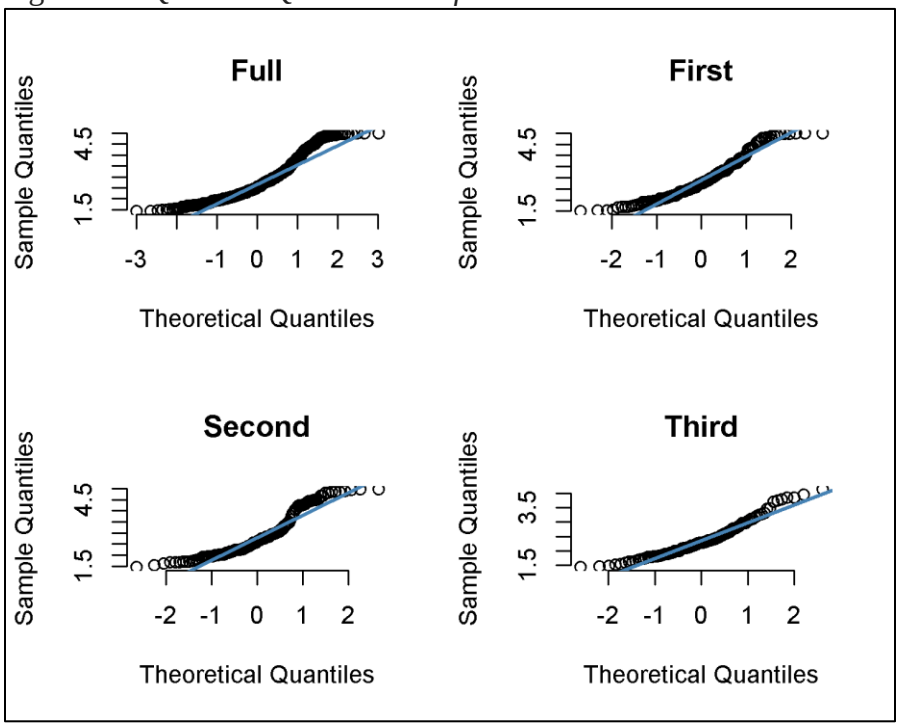


Figure 3.7. *Quantile-Quantile Plot for Phonotactic Probability*

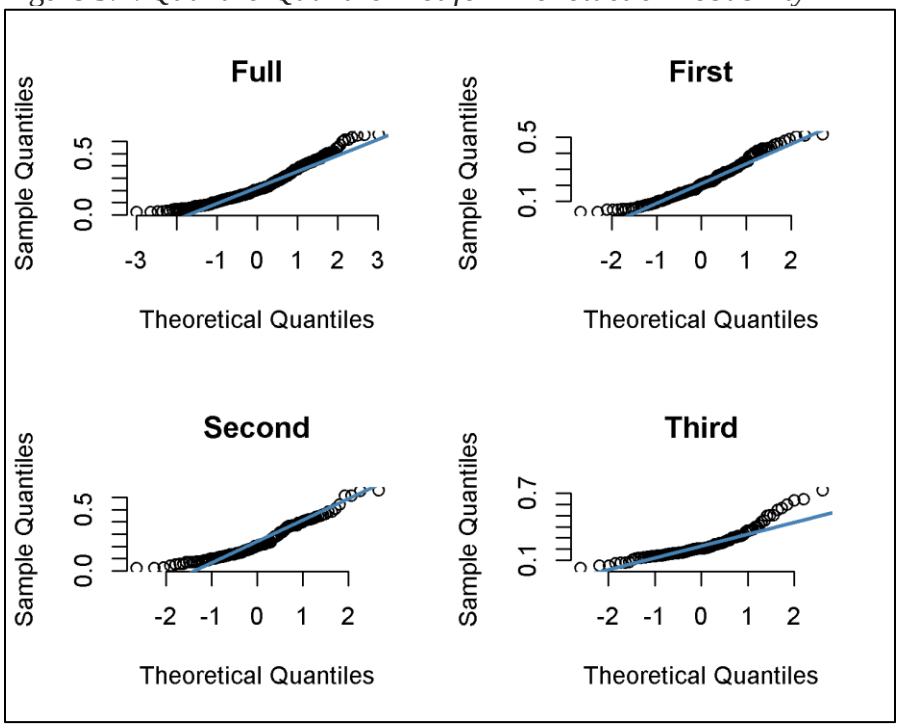
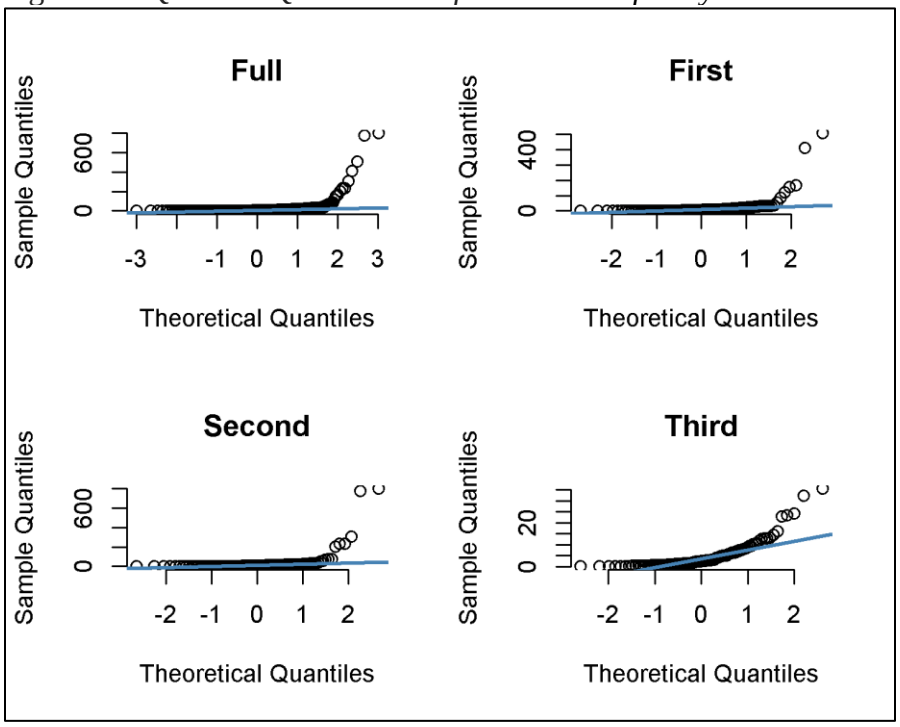


Figure 3.8. *Quantile-Quantile Plot for Word Frequency*



The Shapiro Wilk tested the data for univariate normality, so multivariate normal tests were then conducted to determine if the combined data were normal. Multiple tests were considered for thoroughness. There exist a variety of tests for determining if a dataset is multivariate normal using different techniques. One of the approaches that has been suggested for testing a multivariate normal distribution is to consider skewness and kurtosis. For general multivariate data, Mardia suggested statistics for these measurements (Mardia, 1970). For skewess the test statistic is

$$MS = \frac{1}{6n}\sum_{i,j=1}^{n}\left(Y_i^T Y_j\right)^3$$

and for kurtosis the test statistic is

$$MK = \sqrt{\frac{n}{8p(p+2)}}\left\{\frac{1}{n}\sum_{i=1}^{n} \| Y_i \|^4 - \frac{p(p+2)(n-1)}{n+1}\right\}.$$

The hypothesis of multivariate normality is rejected if skewness MS is too large or if the absolute value of the centralized kurtosis |MK| is large and exceeds a critical value. These tests of simple and informative, providing specific information about the non-normality of the data. One of the drawbacks of this method is that it is not consistent for testing general alternatives and can have low power against many alternatives.

The Doornik-Hansen test expands on Mardia's test for skewness and kurtosis by transforming the multivariate normal to independent standard normal (Doornik and Hansen,

2008). To begin, let $X' = (x_1, \dots, x_n)$ be the $p \times n$ matrix of n observations of p-dimensional vectors. The data has a sample mean

$$\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$$

and the covariance matrix

$$S = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'.$$

With $V = diag\left(\hat{\sigma}_1^2, \dots, \hat{\sigma}_p^2\right)$, the correlation matrix $C = V^{-\frac{1}{2}} S V^{-\frac{1}{2}}$ can be formed and the $p \times n$ matrix $Y' = (y_1, \dots, y_n)$ is defined for the transformed observations:

$$y_i = H\Lambda^{-\frac{1}{2}} H' V^{-\frac{1}{2}} (x_i - \bar{x}).$$

$\Lambda = diag\left(\lambda_1, \dots, \lambda_p\right)$ is the matrix of eigenvalues on the diagonal of C. The columns of H correspond to the eigenvectors such that $H'H = I_p$ and $\Lambda = H'CH$, as well as $n^{-1}Y'Y = I_p$. A multivariate normal can be transformed into independent standard 33ormal using population values of C and V. Sample values can be used to approximate this.

Now the univariate skewness and kurtosis can be computed for each of the transformed vectors by defining $B_1' = \left(\sqrt{b_{11}}, \dots, \sqrt{b_{1p}}\right)$ and $B_2' = \left(b_{21}, \dots, b_{2p}\right)$. The test statistic is

$$E_p^a = \frac{nB_1'B_1}{6} + \frac{n(B_2 - 3_l)'(B_2 - 3_l)}{24} \cong \chi^2(2p)$$

where l is a p-vector of ones. The multivariate statistic is

$$E_p = Z_1'Z_1 + Z_2'Z_2 \cong \chi^2(2p)$$

where $Z_1' = (z_{11}, \ldots, z_{1p})$ and $Z_2' = (z_{21}, \ldots, z_{2p})$. As with Mardia's test, this test is informative about the normality of the multivariate distribution.

The Henze-Zirkler test for multivariate normal is based on the empirical characteristic function and known for having good power and being a consistent test (Henze & Zirkler, 1990). This is based on a nonnegative function that uses characteristic functions to measure the distance between a hypothesized function and an empirical function. For the test to be consistent, the function will equal zero if the data is from a multivariate normal distribution.

This nonnegative function is given by

$$D_\beta(P, Q) = \int_{\Re p} \left|\hat{P}(t) - \hat{Q}(t)\right|^2 \varphi_\beta(t) dt$$

where $\hat{P}(t)$ is the characteristic function of the proposed function and $\hat{Q}(t)$ is the empirical characteristic function being compared with a weighting function $\varphi_\beta(t)$ and the smoothing function $\beta = \frac{1}{\sqrt{2}}\left\{\frac{n(2p+1)}{4}\right\}^{p+4}$ is a smoothing parameter. The test statistic for this is

$$D_{n,\beta} = \frac{1}{n^2}\sum_{j,k=1}^{n} \exp\left(-\frac{\beta^2}{2} \parallel Y_j - Y_k \parallel^2\right) + (1 + 2\beta^2)^{-\frac{p}{2}}$$

$$-\frac{2}{n}(1 + \beta^2)^{-\frac{p}{2}} \sum_{j=1}^{n} \exp\left(-\frac{\beta^2 \parallel Y_j \parallel^2}{\{2(1 + \beta^2)\}}\right).$$

Normality is tested with $HZ_\beta = n\left(4I_E + D_{n,\beta} I_{E^c}\right)$ where $\beta \in R$, $I_E, and\ I_{E^c}$ are indicator functions with $E = \{S_2\ is\ singular\}$ and in terms of Y_i. If HZ_β is too large, multivariate normality is rejected. One shortcoming of this test is that it is not informative about why the data fails the normality test.

Royston's H test is an expansion of the Shapiro-Wilk W univariate normality test for multivariate data (Royston, 1992). To set this up, let W_j denote the value of the jth variable in a p-variate distribution for Shapiro-Wilk W. Next, the test statistics is defined as

$$R_j = \left\{ \Phi^{-1} \left[\frac{1}{2} \Phi \left\{ - \frac{\left(\left(1 - W_j\right)^{\lambda} - \mu \right)}{\sigma} \right\} \right] \right\},$$

with $\lambda, \mu, and\ \sigma$ being calculated from polynomial approximations and $\Phi(\cdot)$ is the standard normal cdf. For multivariate normal data, $H = \xi \sum \frac{R_j}{p}$ is approximately $\chi^2_{\hat{\xi}}$ distributed with $\hat{\xi} = \frac{p}{[1+(p-1)\bar{c}]}$ and $\bar{c}$ is an estimate of the average correlation among $R_j's$. The $\chi^2_{\hat{\xi}}$ distribution is used to obtain the critical values for this test. The Royston's H test for multivariate normality has good power against many alternative distributions.

The energy test for multivariate normal is a goodness of fit test that has been shown to often outperform common tests such as Mardia's test or Henze-Zirkler (Székely and Rizzo, 2005). To begin, suppose that $X_1, \dots, X_n$ be a random sample from a d-variate population with distribution F and $x_1, \dots, x_n$ are the observed values of the random sample. The test statistic testing $H_0: F = F_0\ vs\ H_1: F \neq F_0$ is

$$\mathcal{E}_{n,d} = n \left(\frac{2}{n} \sum_{j=1}^{n} E \parallel y_j - Z \parallel - E \parallel X - X' \parallel - \frac{1}{n^2} \sum_{j,k=1}^{n} \parallel x_j - x_k \parallel \right),$$

where X and X' are independent and identically distributed with distribution F_0.

Table 3.7 has the results for Mardia skewness and kurtosis, the Doornik-Hansen test, Henze-Zirkler test, Royston test, and energy E test for multivariate normality for each data subset. These were calculated in R using the mvn package (Korkmaz et al,). Each of the results is

significant, meaning that every subset was found to be non-normal by every chosen test. Each test rejected the null hypothesis that the data follows a normal distribution.

Table 3.7. *Multivariate Normality*

	Full Model	First Grade	Second Grade	Third Grade
Mardia Skewness	9050.76*	2130.57*	1869.63*	2063.24*
Mardia Kurtosis	166.92*	62.00*	74.20*	65.53*
Doornik-Hansen	8031.62*	1751.04*	1389.05*	907.94*
Henze-Zirkler	9.55*	7.27*	6.31*	4.37*
Royston	1.14*	231.68*	235.93*	191.65*
Energy	475783835104*	11.87*	10.43*	2484.60*

Because of the non-normal nature of the data, a nonparametric ranked correlation test might be appropriate (Conover, 1998). The two rank correlation tests considered for this are Spearman's rho and Kendall's Tau. For Spearman's rho the data is ranked from 1 to n, with 1 being the smallest data point and n being the largest, or $R(X_i) = 1$ if X_i is the smallest. The measurement of correlation is given by

$$\rho = \frac{\sum_{i=1}^{n} R(X_i)R(Y_i) - n\left(\frac{n+1}{2}\right)^2}{\left(\sum_{i=1}^{n} R(X_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)^{\frac{1}{2}}\left(\sum_{i=1}^{n} R(Y_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)^{\frac{1}{2}}}.$$

The results for Spearman's rho ranked correlation test are found in Table 3.8, for the full dataset, first grade, second grade, and third grade. Both ranked correlation tests are included as part of base R (R Core Team, 2019).

Kendall's tau compares two observations and is concordant if both higher values come from the same observation. Let N_c be the number of concordant pairs of observations and N_d be the number of discordant pairs when the two numbers of on observation differ in direction. There are $\binom{n}{2} = \frac{n(n-1)}{2}$ total possible pairs which accounts for pairs N_c, N_d, and ties. The test statistic for Kendall's tau is

$$\tau = \frac{N_c - N_d}{n(n-1)/2}.$$

If all observations are concordant $\tau = 1$ and if all observations are discordant $\tau = -1$. The results for Kendall's tau ranked correlation test are found in Table 3.9 for the full dataset, first grade, second grade, and third grade.

Table 3.8. *Spearman's Rho Ranked Correlation*

Full Model	W Freq	AoA	Concrete	Nden	Phon Prob
W Freq	1				
AoA	-.62*	1			
Concrete	.25*	-.44*	1		
Nden	.49*	-.46*	.28*	1	
Phon Prob	-.10	.21*	-.16*	-.30*	1
First Grade					
W Freq	1				
AoA	-.52*	1			
Concrete	.32*	-.56*	1		
Nden	.41*	-.41*	.36*	1	
Phon Prob	-.03	.31*	-.18*	-.37*	1
Second Grade					
W Freq	1				
AoA	-.65*	1			
Concrete	.16	-.44*	1		
Nden	.54*	-.44*	.23*	1	
Phon Prob	-.18*	.14	-.18*	-.30*	1
Third Grade					
W Freq	1				
AoA	-.54*	1			
Concrete	-.03	.00	1		
Nden	.39*	-.42*	.06	1	
Phon Prob	.07	.21*	-.07	-.25*	1

Both Spearman's and Kendall's ranked correlation tests agree with the previous tests. Every dataset has moderate correlation, mutual information, multicollinearity, and ranked correlation among the variables. While performing these tests, it was also determined that the data does not follow univariate normal distributions for the variables and is not normally distributed for the multivariate data. Based on the exploratory data analysis, some basic linear models may not be applicable. Advanced statistical learning and machine learning can be used to account for these data conditions.

Table 3.9. *Kendall's Tau Ranked Correlation*

Full Model	W Freq	AoA	Concrete	Nden	Phon Prob
W Freq	1				
AoA	-.45*	1			
Concrete	.16*	-.31*	1		
Nden	.35*	-.32*	.19*	1	
Phon Prob	-.05	.14*	-.11*	-.21*	1
First Grade					
W Freq	1				
AoA	-.37*	1			
Concrete	.21*	-.40*	1		
Nden	.28*	-.29*	.26*	1	
Phon Prob	-.02	.21*	-.11*	-.26*	1
Second Grade					
W Freq	1				
AoA	-.48*	1			
Concrete	.10	-.31*	1		
Nden	.39*	-.31*	.16*	1	
Phon Prob	-.13*	.09	-.11	-.20*	1
Third Grade					
W Freq	1				
AoA	-.40*	1			
Concrete	-.02	.00	1		
NDen	.27*	-.30*	.04	1	
Phon Prob	.58	.14*	-.05	-.16*	1

CHAPTER FOUR:

STATISTICAL METHODS USED TO ANALYZE WORD LEARNING DATA

Introduction

Investigators have examined the effects of lexical characteristics, such as word frequency, age of acquisition (AoA), neighborhood density, and phonological phonotactic probability on word learning in children using both multivariate linear regression and stepwise regression (Gray, 2004; Morrison & Ellis, 2000; Stoel-Gammon, 2010; Storkel, 2009). For example, Morrison and Ellis (2000) used multivariate linear regression to examine factors that impacted children's word naming speed and found that AoA, word length, word frequency, and orthographic neighborhood density significantly predicted naming speeds.

Multivariate linear regression and stepwise regression models are well-known by researchers and are easy to implement and interpret. However, they rely on certain assumptions, and if these assumptions are not satisfied results can be unreliable (Osborne & Waters, 2002). Multivariate linear regression models assume that there is a linear relationship between the dependent and independent variables. It assumes that independent variables do not interact with each other, that the data is normally distributed, and that the residuals are normally distributed with a mean of 0 and common variance of σ (homoscedasticity). When many of these assumptions are not met its applicability is limited and it may not account for ambiguous data such as unplanned data, latent variables, or correlation (Draper & Smith, 1998).

Stepwise regression is similar to multivariate linear regression; it is based on the same foundations and relies on the same assumptions. But stepwise regression differentiates itself

from multivariate linear regression by performing variable selection that is faster than other auto-selection models. This procedure begins with an empty model and may add one variable at each step or begins with a full model and may delete one variable at each step. Variable selection will fine tune the model by adding or removing variables, until it has chosen the best predictor variables while excluding insignificant ones. This provides valuable information based on the order the variables are added or removed. However, the most accurate model may not be chosen because it does not consider every possible combination of variables because it uses greedy selection criteria. The variable selection process of stepwise regression requires a larger sample size and is sensitive to multicollinearity and redundant predictors.

Resulting multiple and stepwise regression models are easy to interpret. Both provide a regression plane that represents a positive, negative, or neutral trend. Higher dimensional trends are difficult to visualize but individual trend lines can be viewed by holding all other variable constant, but again only the trend of the data is displayed. This can lead to a generalized description of the data, but finer details could be lost. For example, age of acquisition is related to word learning; words learned earlier are easier to use than those learned later (Kupperman et al., 2012), but this association may not be a one-to-one linear relationship. There could be periods of rapid lexical growth followed by prolonged periods of growth (Goldfield & Reznick, 1990). Regression models would show this growth as a steady incline but would not differentiate between these accelerated or slowed periods of learning. Valuable information about when and how children learn words would be missed.

Additionally, insignificant variables may remain in the models leading to inflated R^2 values compared to the model without the insignificant variables (Tamhane & Dunlop, 2000). Linear models offer simplicity at the cost of making numerous assumptions about the data that

may not be accurate. More advanced models using statistical learning and machine learning may offer stronger results while being more reliable based on the conditions of the data.

When determining which statistical methods are appropriate for analysis, it is important to consider the data being analyzed. Exploratory data analysis was performed for the data to determine which models would best describe the data. The ILIAD data was found to be nonlinear, have moderate multicollinearity, skewed parameters, the variables were not univariate normal, the overall data was not multivariate normal, and had variable heteroscedasticity. Each of these conditions can cause difficulty with many models, so it is vital to choose a model that limits reliance on assumptions about these conditions.

When selecting analytic techniques, it is important to consider the level of expertise needed to implement and interpret results. Model selection should balance choosing an appropriate model given the aspects of the ILIAD data with the ability to interpret results to provide meaningful outcomes relevant to educational researchers. There is no need to use complex models for fitting data when simpler, linear techniques are adequate.

Models were selected to get progressively more complex while addressing conditions of the data. Linear shrinkage methods were selected to handle the multicollinearity of the data. While they rely on many of the same assumptions as simpler models, such as linearity and homoscedasticity, their ability to deal with multicollinearity is an intermediate step between v and those that are more complex. The data has moderate variable collinearity, including ranked correlations, and prior studies have indicated a connection between lexical characteristics, to justify the consideration of shrinkage methods (Hoover et al, 2010; Storkel, 2004; Vitevitch et al, 2004).

Discrete variable selection methods, such as stepwise regression, retain a subset of the predictors and discard the rest. This subset selection can model the data with lower prediction error than the full model and may lead to a more interpretable model but because of its discrete nature, the variance is often higher than continuous variable selection methods. This is caused by the "all-or-nothing" approach of deleting variables for discrete variable selection so more continuous methods do not have such high variability. Shrinkage methods are more continuous and can decrease variability. The shrinkage methods included in our analysis are ridge regression (Hoerl & Kennard, 1970), least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), principal components regression (PCR; Massy, 1965), and partial least squares (PLS; Wold, 1975). These models were chosen because of their ability to deal with multicollinearity, which the ILIAD dataset exhibits.

Shrinkage methods of regression are penalized models with the aim to lessen variance. These methods tune the parameters to ordinary least squares, keeping the model interpretable. The tuning parameters for ridge regression and LASSO vary over a continuous range while PLS and PCS tune to the best subset in discrete steps. Ridge regression shrinks in all directions but shrinks low variance directions more than higher variance directions never eliminates variables (Friedman et al, 2001). LASSO can eliminate variables, as can PCR which leaves M high variance directions alone and ignores the rest. PLS, like ridge regression, shrinks low-variance directions but can also inflate some higher variance directions. PLS can be viewed as a supervised alternative to PCR (James et al, 2013) that seeks directions with high variance and high correlation while PCR focuses only on high variance (Stone & Brooks, 1990). Ridge regression, PLS, and PCR often behave similarly with LASSO as an intermediate, but ridge regression is preferred for minimizing prediction error because it shrinks smoothly (Frank &

Friedman, 1993). Generally, shrinkage methods are considered weak for skewed data, but researchers have shown that ridge regression and LASSO can do well with skewed data and partial least squares is robust to skewed data (Cao et al, 2021; Shutes & Adcock, 2013; Cassel et al, 1999). Elastic net regularization and variable selection is a method that often outperforms LASSO while enjoying similar sparsity of representation (Zou & Hastie, 2005). This is done by combining the L_1 and L_2 penalties from LASSO and ridge regression, respectively (Mol et al, 2009). In general, the accuracy and bias for elastic nets falls between ridge regression and LASSO. Empirically, ridge regression and LASSO regression performed equally for the ILIAD data, so it was unlikely that elastic nets would offer improvement to regressing the data. Because of this, elastic nets were not included in the comparison of shrinkage methods.

Support vector regression (SVR) is a robust machine learning model that relies on very few assumptions. SVR is adaptable based on kernel selection, which must be determined for the data being considered. The selection of different kernels will determine the strength of this model. Support vector regression was chosen because of its power and flexibility, as well as its ability to work with nonlinear data. Support vector regression splits the data with a hyperplane to split the data and find the best fit using support vectors. The kernels that determine how the hyperplane is created give SVR the capability to handle nonlinear data. Choosing the correct kernel is very important and often requires some prior knowledge about the data or thorough exploratory data analysis. The choice of kernel and inclusion of a cost function can make support vector regressions difficult without some expertise.

Neural networks were considered for fitting the data, but prior studies have shown that SVR can often match it (Were et al, 2015; Cortes & Vapnik, 1995). Neural networks can approximate real-valued functions (Song et al, 2017) require more data as their grown in

complexity, which would be necessary for the ILIAD data (Hagan et al, 1997). With the limited data within grade levels, this would lead to questionable results. With this consideration, neural networks were not included in the analysis for the ILIAD word learning data. Additionally, in order to create a robust, adaptable neural networks, researchers would need a level of expertise that limits its usefulness in the field of education, a priori knowledge and optimally tuned parameters would be required (Smola & Schölkopf, 2004), and the models are computationally complex (Šima & Orponen, 2003).

Ensemble tree-based models were chosen because they are robust to most data and have strong predictive performance (Caruana & Niculescu-Mizil, 2006). We decided to focus on tree-based methods such as regression trees, random forest, gradient boosting machines, and stochastic gradient boosting machines. Random forest and gradient boosting are considered to be very strong "out-of-the-box" or "off-the-shelf" models because of their predictive performance, relatively little hyperparameter tuning, and few assumptions (Boehmke & Greenwell, 2019). Regression trees were included because of the fundamental nature of tree building being a part of random forest. Random forests are a tree-based ensemble method for modeling that performs bagging, or bootstrapping the sample data, and then creates numerous non-pruned trees. These trees are creates with binary splits where a subset of the parameters is randomly chosen. This cut point is chosen to reduce RSS when splitting the feature space. The random subset of parameters inputs randomness which decorrelates the individual trees. Gradient boosting is an iterative tree building method that uses the results of a fitted model to update the next recursively until some threshold is met, usually a set number of iterations. Stochastic gradient boosting machines are the same as gradient boosting machines but where a random subset is chosen for each iteration to introduce randomness.

Model Descriptions

Ridge Regression

Ridge regression is a method for estimating the coefficients for multivariate linear regression using shrinkage. It does this by imposing a penalty on the size of the coefficients. Unlike stepwise regression, ridge regression shrinks parameter estimates towards 0 but never hits 0, which would eliminate the variable from the model. Ridge regression does not perform feature selection, though the estimates may become negligibly small (Kuhn & Johnson, 2013).

For multivariate linear regression, the estimates for $\beta_1, \beta_2, \ldots., \beta_p$ are found by minimizing

$$RSS = \sum_{i=1}^{n}\left(y_i - \beta_o - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

in the least squares fitting procedure. Similarly, ridge regression performs the same step with an included penalty. The coefficients are found by minimizing a penalized sum of squares,

$$\hat{\beta}^{ridge} = armin_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \beta_o - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\}$$

where $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage (Friedman et al., 2001). This is because $\lambda\sum_j \beta_j^2$, the shrinkage penalty, is small when $\beta_1, \beta_1, \ldots., \beta_p$ approach zero. The shrinkage penalty is only applied to $\beta_1, \beta_1, \ldots., \beta_p$ to shrink the associated variables with respect to the dependent variable and not the intercept. It is assumed that the variables have been standardized, that is, centered to have a mean of 0.

The complexity parameter λ controls the relative impact of the coefficients, shrinking the coefficients as λ increases. When $\lambda = 0$, the shrinkage penalty has no impact and it is ordinary

least squares. As $\lambda \to \infty$ the coefficient estimates will approach 0, as the shrinkage penalty grows (James et al., 2013). Because this is reliant on the complexity parameter, there exists different coefficient estimates $\hat{\beta}_{\lambda}^{ridge}$ for each value of λ. It is important to find a good value for λ to produce the best mode but choosing an appropriate λ is one of the main challenges with using ridge regression. This is done for a bias and variance tradeoff. By inputting a small amount of bias as λ increases, a larger amount of variance can be removed.

When ridge regression was created, Hoerl and Kennard (1970) suggested using a graphic called the ridge trace. A ridge trace has the standardized betas on the vertical axis and λ levels on the horizontal axis. This allows the researcher to determine how the changes in λ effect the coefficients for each parameter. Similarly, the vertical axis of the ridge trace can be replaced with other metrics such as VIF. The smallest ridge constant where the coefficients have stabilized is preferred. By limiting the size of the ridge constant, the amount of bias introduced should be small. Because this approach is graphical in nature, the results will rely on the expertise of the researcher. The basic graphical approach also suffers from only showing bias directly and ignores the multidimensional nature of the problem (Friendly, 2013).

It is common for researcher to choose a value that is too large using the ridge trace, so Hoerl and Kennard proposed an analytic method for choosing λ (Hoerl and Kennard, 1976). This approach does not necessarily converge, so other methods may be preferable such as using singular value decomposition (Bair et al, 2006). Ridge regression can be a considered a “smooth case” of principal component regression

The assumptions of ridge regression are similar to that of linear regression: linearity, constant variance, and independence. One of the main strengths of ridge regression is the tradeoff between variance and bias. When a regression model has many correlated variables, the

resulting coefficients may have high variance and be poorly selected. As one variable changes, others will likewise change leading to an unstable model that fluctuates significantly given small changes. With shrinkage methods, a large coefficient for one variable can be offset or cancelled out with a similarly small coefficient. The shrinkage of the model will also prevent overfitting.

The increase in bias that is introduced is one of the shortcomings of ridge regression. Because of the shrinkage of the model, it may be hard to interpret the results when compared to simpler regression methods. This is not a problem for the ILIAD data, since the number of variables is not overly large. The model is completely reliant on the best complexity parameter being chosen. Ridge tracing can lead to an incorrect choice because it relies on expertise and analytic selection methods may rely on assumptions about the prior distribution. There are many papers claiming that ridge regression estimates are better than least squares estimates when judged using mean square error but these should be viewed with caution (Draper &Smith, 1998).

Least Absolute Shrinkage and Selection Operator (LASSO)

Generally, there are two reasons that ordinary least square estimates are inadequate, prediction accuracy and interpretation. OLS often has low bias but large variance, where shrinking or setting some of the coefficients to zero can lower variance but introduce a small amount of bias, therefore increasing the prediction accuracy. If there exists a large number of predictor variables, interpretation may be unfeasible unless a smaller subset is selected that accounts for most of the effects. This is exactly what the least absolute shrinkage and selection operator (LASSO) was introduced to do (Tibshirani, 1995). While ridge regression shrinks coefficients as well and is very stable, it is unable to set them to 0, which LASSO can. LASSO attempts to keep the strengths of both subset selection and ridge regression.

Like ridge regression, LASSO solves an ordinary least squares with a penalty. The subtle difference is the penalty term the model is subjected to. LASSO is defined by

$$\hat{\beta}^{lasso} = argmin_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

$$subject\ to \sum_{j=1}^{p} |\beta_j| \leq t,$$

or equivalently

$$\hat{\beta}^{lasso} = argmin_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

As with ridge regression, β_0 is reparametrized by standardizing the predictors. With this formulation, if t is sufficiently small some coefficients will shrink to 0. This penalty will, increases (or decreases) the coefficient of the first variable only as long as its correlation with the residual is larger than that of the inactive predictors (Lockhart et al, 2014).” Subsequent steps will follow the same way iteratively.

LASSO has the same assumptions for modeling the data as ridge regression. Its ability to shrink the data and eliminate variables can create a simpler, more interpretable model. This shrinkage will fitting. By limiting the number of parameters included in the model, they will be more biased than before shrinkage. If there is a group of correlated features, LASSO will select the ones to keep arbitrarily. The performance is generally slightly worse that ridge regression as a result of the simpler model.

There is no direct way to test the significance for the coefficients chosen by LASSO but “The lasso, on the other hand, increases (or decreases) the coefficient of the first variable only as long as its correlation with the residual is larger than that of the inactive predictors” (Lockhart et

al, 2014, p. 34). The major advantage of LASSO over ridge regression is that it can produce simpler, more interpretable models that only rely on a subset of the variables, but there is no universal "better model" between the two (James et al, 2013).

Partial Least Squares (PLS)

It is often the case with real world data like ILIAD, predictors can be correlated and contain similar predictive information. If there is a high amount of correlation among the predictor variables, then ordinary least squares for multivariate linear regression will become unstable (Kuhn & Johnson, 2013), meaning it may not be possible to find a unique set of regression coefficients. PLS can be viewed as a supervised method similar to principal component regression, which is unsupervised, that focuses on the directions with highest variance and high correlation, while PCR only focuses on high-variance directions. Ordinary least squares may also be unable to find a unique set of regression coefficients if the number of predictors is higher than the number of observations. The underlying assumption of partial least squares is that data is generated by a process that is driven by a smaller number of latent factors (Wold, 1975).

Partial least squares (PLS) is a supervised method of dimensional reduction, that is, it uses y in addition to X during creation. Partial least squares is not scale invariant, meaning it is assumed that x_j is standardized with a mean of 0 and variance of 1 (Friedman et al, 2001). The first step of PLS is to identify a new set of features $Z_1, \dots, Z_M$ that are linear combinations of the original features $X_1, \dots, X_p$. A linear regression model is then fit with these M new features.

The first step is to standardize the variables and compute $\hat{\varphi}_{1j} = \langle X_j, Y \rangle$ for each j, that is, to compute the first direction Z_1 by setting each $\hat{\varphi}_{1j}$ equal to the coefficient from a linear regression of Y onto X_j. The coefficient will be proportional to the correlation between Y and X_j

which means the inputs have a weight associated with their univariate effect on Y. Deriving the first partial least squares direction $Z_1 = \sum_j \hat{\varphi}_{1j} X_j$ puts the highest weight on variables that are most strongly associated with the response.

Next each variable will be adjusted for Z_1 by regressing each variable onto it, getting $\hat{\theta}_1$ and finding the residuals, representing the amount of information not explained by the first PLS direction. The second PLS direction Z_2 is computed using this orthogonalized data and completing the same steps as for Z_1. This process is continued until $M \leq p$ directions have been obtained, where M is obtained using cross validation. If $M = p$ directions, PLS would be the same as ordinary least squares, while $M < p$ produces a regression that has been reduced. This process is described in table 4.1.

Partial least squares is an extension of multivariate linear regression, so it has many of the same underlying assumptions but they are generally not as concrete. The method struggles with outliers and nonlinear data relationships. There is an assumption of some underlying system that the latent variables represent, but no particular distribution is assumed. This makes testing the significance of the resulting coefficients difficult without other steps such as bootstrapping. Even with bootstrapping, the results may not be reliable with a small sample size as it is likely to fit the noise instead of the true distribution.

It is able to model multiple dependent and independent variables, called PLS1 for single dependent variable and PLS2 for multiple (Rosipal & Krämer, 2005). One of the major strengths of it as a shrinkage method is that it can handle multicollinearity in the independent variables very well and it is robust to noise and missing data. The original method has been expanded with various techniques to allow for modeling nonlinear relationships without the loss of interpretability (Rosipal, 2011). By focusing on latent variables with the highest impact, it often

leads to stronger predictions. The model is distribution free and can handle a variety of variables (categorical, ordinal, interval).

Table 4.1. *Partial Least Squares Algorithm* (Friedman et al, 2001)

1. Standardize each x_j to have mean 0 and variance 1. Set $\hat{y}^{(0)} = \bar{y}\mathbf{1}$ and $x_j{}^{(0)} = x_j$, $j = 1, \ldots, p$.
2. For $m = 1, \ldots, p$
   a. $z_m = \sum_{j=1}^{p} \hat{\varphi}_{mj} x_j{}^{(m-1)}$, where $\hat{\varphi}_{mj} = \langle x_j{}^{(m-1)}, y \rangle$.
   b. $\hat{\theta}_m = \langle z_m, y \rangle / \langle z_m, z_m \rangle$.
   c. $\hat{y}^{(m)} = \hat{y}^{(m-1)} + \hat{\theta}_m z_m$.
   d. Orthogonalize each $x_j{}^{(m-1)}$ with respect to z_m: $x_j^{(m)} = x_j^{(m-1)} - \left[\langle z_m, x_j^{(m-1)} \rangle / \langle z_m, z_m \rangle\right] z_m, j = 1, \ldots, p$.
3. Output the sequence of fitted vectors $\{\hat{y}^{(m)}\}_1^p$. Since the $\{z_\ell\}_1^m$ are linear in the original x_j, so is the $\hat{y}^{(m)} = \boldsymbol{X}\hat{\beta}^{pls}(m)$. These linear coefficients can be recovered from the sequence of PLS transformations.

Partial least squares works well on small datasets, though it makes it difficult to test the significance without knowledge of the underlying distribution. One of the biggest weaknesses for the method is the lack of model test statistics. It can be difficult to interpret the loadings of independent latent variables.

Partial least squares is a discrete least squares solution that can be a little bit unstable, causing it to have slightly higher prediction error compared to ridge regression (Frank & Friedman, 1993). PLS is and PCR both roughly track the ridge path but their discrete nature makes them more extreme (Friedman et al, 2001). Principal components regression leaves M high-variance directions alone and discards the rest while PLS focuses on high-variance directions with high correlation.

Principal Component Regression (PCR)

Principal component regression was originally formulated by Massy to bring together the ideas of regression and principal component analysis (Massy, 1965). He argued that by

transforming a set of variates into principal components, their relations with another variable may be explored more easily. If the independent variables are highly collinear or there are a large number of predictors, it may be beneficial to simplify the sample space to principal components. The dependent variable could then be regressed on the resulting principal components. It is an unsupervised machine learning model but can be altered to be semi-supervised or supervised (Bair et al, 2006).

In principal component regression (PCR), the first $M \leq p$ principal components $Z_1, \dots, Z_M$ are constructed and these are used as the predictors in a linear regression fit with ordinary least squares. The derived input columns $z_m = Xv_m$ and then y is regressed on $z_1, \dots, z_M$. This will form a sum of univariate regressions

$$\hat{y}_{(M)}^{pcr} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m z_m,$$

where $\hat{\theta}_m = \langle z_m, y \rangle / \langle z_m, z_m \rangle$, since the z_m are orthogonal. The solution to this can be expressed in terms of the coefficients of the original x_j since the z_m are linear combinations of them.

$$\hat{\beta}^{pcr}(M) = \sum_{m=1}^{M} \hat{\theta}_m v_m$$

Principal components depend on scaling the inputs as with ridge regression. In fact, ridge regression and PCR are both scaling operations that operate based on principal components but where ridge regression shrinks the coefficients for principal components, PCR discards the $p - M$ smallest components.

This approach assumes that the directions in which $X_1, \dots, X_p$ show the most variation are the directions most associated with Y. This is not always true but generally is correct enough to justify the assumption (James et al, 2013). If this assumption is true, then the least squares model

fitted to $Z_1, \dots, Z_M$ will be better than the least squares model fitted to $X_1, \dots, X_p$. This is because most of the data is contained in $Z_1, \dots, Z_M$ and will prevent overfitting.

The model contains some of the assumptions that are the same as with regular multivariate linear regression. PCR assumes the data is linear, there is constant variance, there are no outliers, and independence. The model does not assume that the data is normally distributed. It is assumed that the latent features are independent and identically distributed (i.i.d) and that many large singular values do not exist (Agarwal et al, 2021). Because the ILIAD data does not contain a large number of variables, there will not be a large number of latent variables found. While shrinking the model, it will minimize overfitting and may make it easier to visualize by lowering high dimensional data. This can lead to a loss in interpretability and there will be some amount of information loss.

Support Vector Regression (SVR)

Support vector machines are a powerful, flexible model that was originally created to work for classification problems. The goal of support vector machines is to find a hyperplane to separate the classes by maximizing the margin between the groups. The points on these margins are known as support vectors. If the hyperplane can perfectly separate the classes, then an infinite number of such hyperplanes will exist. To choose the best hyperplane, using the maximal margin hyperplane is an option that is furthest from the training observations by computing the perpendicular distance of each training observation from a given hyperplane, known as the margin.

To adapt the support vector machines from classification, consider the linear regression model

$$f(x) = x^T\beta + \beta_0.$$

In order to handle nonlinear data, we estimate β by minimizing

$$H(\beta, \beta_0) = \sum_{i=1}^{N} V(y_i - f(x_i)) + \frac{\lambda}{2}\|\beta\|^2,$$

where

$$V_\epsilon(r) = \begin{cases} 0 & if\ |r| < \epsilon, \\ |r| - \epsilon, & otherwise. \end{cases}$$

Here r is the distance from the output to the target, λ is the regularization parameter, and ϵ is the error. This method uses an "ϵ-insensitive" error measure that assumes under the conditions where y is the result of a function with normal additive noise, it will be the best approximation for the regression (Vapnik, 2013). The algorithm can be modified so that ϵ does not need to be specified a priori by specifying an upper bound ν (Smola & Schölkopf, 1998). If the additive noise is from another distribution, another optimal approximation may be more appropriate to use.

One of the drawbacks of minimizing sum of square errors is that the parameter estimates can be influenced by just one observation which is far away from the general trend. A more robust error measure to use is a loss function that only depends on the density describing the noise (Huber, 1964). This takes the form

$$V_H = \begin{cases} \frac{r^2}{2} & if\ |r| \le c, \\ c|r| - \frac{c^2}{2} & |r| > 2. \end{cases}$$

This is a function similar to a Huber function where the contributions from observations are reduced from quadratic to linear. This is done using the absolute residual greater than some constant c that is chosen beforehand (Friedman et al, 2001). This causes fitting to be less

sensitive to outliers by not using squared residuals and samples with small residuals no effect on the regression equation (Kuhn & Johnson, 2013).

The function will have the form

$$\hat{\beta} = \sum_{i=1}^{N} (\hat{\alpha}_i^* - \hat{\alpha}_i) x_i,$$

and

$$\hat{f}(x) = \sum_{i=1}^{N} (\hat{\alpha}_i^* - \hat{\alpha}_i) \langle x, x_i \rangle + \beta_0,$$

if $\hat{\beta}$ and $\hat{\beta}_0$ are minimizers of H. Here $\hat{\alpha}_i^* \; and \; \hat{\alpha}_i$ are positive values that solve the quadratic programming problem

$$min_{\hat{\alpha}_i^*, \hat{\alpha}_i} \epsilon \sum_{i=1}^{N} (\hat{\alpha}_i^* - \hat{\alpha}_i) + \frac{1}{2} \sum_{i=1}^{N} y_i (\hat{\alpha}_i^* - \hat{\alpha}_i) + \frac{1}{2} \sum_{i,i'=1}^{N} (\hat{\alpha}_i^* - \hat{\alpha}_i)(\hat{\alpha}_{i'}^* - \hat{\alpha}_{i'}) \langle x_i, x_{i'} \rangle.$$

This is subject to the constraints

$$0 \leq \alpha_i, \alpha_i^* \leq 1/\lambda$$

$$\sum_{i=1}^{N} (\hat{\alpha}_i^* - \hat{\alpha}_i) = 0$$

$$\alpha_i \alpha_i^* = 0.$$

The solution values for $(\hat{\alpha}_i^* - \hat{\alpha}_i)$ will generally only be nonzero for a subset (Bishop, 2006). The associated data values are called the support vectors and the solution only depends on the inner products $\langle x_i, x_{i'} \rangle$. This is where the choice of kernels determines the outcome of the model. The kernel function

$$K(x, x') = \langle h(x), h(x') \rangle$$

computes the inner products in the transformations $h(x)$ and $h(x')$. By changing the convolution of the dot product, we can implement different networks (Cortes & Vapnik, 1995). Many kernels exist to choose from, such as linear, Gaussian, polynomial, radial and sigmoid. Table 4.2 contains some commonly used kernels that may be used for support vectors but is not exhaustive.

Table 4.2. *Kernels (*Smola & Schölkopf, 1998*)*

Kernel	Transformation
Linear	$\langle x, x' \rangle$
Polynomial	$\langle x, x' \rangle^d$
Gaussian	$\exp\left(\frac{\|x - x'\|^2}{2\sigma^2}\right)$
Sigmoid	$\tanh(\kappa\langle x, x' \rangle + \Theta)$
Radial Basis Function	$f\big(d(x, x')\big)$

Note. $d \in \mathbb{N}$ and $\sigma, \kappa, \Theta \in \mathbb{R}$, d is a metric on χ and f is a function on $\mathbb{R}$.

When solving the quadrative programming problem, a loss function or cost function can give better results by being a regularization parameter, especially when the outputs are not clearly split. This regularization parameter C or λ is a common difficulty for many practitioners (Hastie et al, 2004). Many software packages will have default C values that are often used without any further consideration. Hastie, Rosset, Tibshirani, and Zhu (2004) were able to develop a method for testing each value of the cost function while maintaining computational complexity. The loss functions are part of the flexibility of support vector regression. Loss functions can be combined into a local loss function, and one may define it pointwise for each sampling point (Smola, 1996).

Support vector regression is a strong, flexible method for modeling data, especially when there are higher dimensional spaces. It excels when the number of dimensions is greater than the

sample size and when there are clear margins of separation in the outputs. It can, however, overfit the data is p is large and the hyperplane is sensitive to single observations that cannot be separated (outliers), which may lead to overfitting.

Regression Trees

Regression trees are a part of classification and regression trees (CART) which partition or stratify the feature space into a set of simple regions and then fits a simple model in each one (Friedman et al, 2001). Trees are created that start with a split in one of the predictors called a node that branches to further nodes called internal nodes. At each internal node, another split occurs for one of the predictors that is based on which split of the feature space causes the greatest decrease in RSS. The branches end at a terminal node or leaves of the tree. Generally, the splits closer to the base of the tree are considered the most important.

The regions can have any shape, but for a simple case consider dividing the space into high-dimensional rectangles (boxes). Boxes $R_1, R_2, \ldots, R_j$ need to be found that minimize the residual sum of squares (RSS),

$$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j} \right)^2$$

where $\hat{y}_{R_j}$ is the mean response for the training observations within the jth box. Calculating every possible partition of the feature space is computationally infeasible so a greedy approach or recursive binary splitting is used. The greedy algorithm is a top down because it begins with a single region with all possible observations and then a split occurs. Each region is successively split until some minimum number of observations at each terminal node. The approach is called greedy because it looks for the best solution for the current step, with no consideration for future steps that may lead to a better overall tree.

Binary recursive splitting selects the predictor X_j and a cut point s that splits the feature space into two regions, $\{X|X_j < s\}$ and $\{X|X_j \geq s\}$, that has the greatest decrease in RSS. For any j and s, two half-planes will take the form

$$R_1(j,s) = \{X|X_j < s\} \; and \; R_2(j,s) = \{X|X_j \geq s\}$$

where j and s are chosen to minimize

$$\sum_{i:x_i \epsilon R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_i \epsilon R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2.$$

For the training observations in $R_1(j,s)$, $\hat{y}_{R_1}$ is the mean response $\hat{y}_{R_2}$ is the mean response for the training observations in $R_2(j,s)$. This is performed iteratively for each region and continues until some ending criterion is met.

Generally, the resulting tree will be too large and complex which leads to overfitting. A smaller tree, with fewer splits may have lower variance and better interpretations with minimal input of bias. There are different approaches to accomplish this. Having a high threshold for minimizing RSS when splitting the feature space is one option. This is a shortsighted approach though because a relatively weak split may lead to a significantly better split afterwards. Instead pruning can be used to grow a large tree, T_0, and then prune it back to some subtree with the lowest test error rate.

One of the options for doing this is cost complexity pruning, or weakest link pruning. This is done by considering a sequence of trees indexed by a nonnegative tuning parameter α, rather than every subtree. For each value of α there will be a subtree $T \subset T_0$ such that

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2 + \alpha|T|.$$

|T| is the number of terminal nodes in tree T, R_m is the subset of predictor space corresponding to the mth terminal node, and $\hat{y}_{R_m}$ is the predicted response associated with R_m. The tuning parameter α controls the tradeoff between a subtree's complexity and the fit of the data. As α increases, the there is a cost for having more terminal nodes on a tree. When $\alpha = 0$, the subtree of T is T_0 and branches will be pruned in a predictable, nested way as α increases from 0. The algorithm for building a regression tree can be found in Table 4.3.

Regression trees are robust and flexible models that do not require standardization or normalization. It is easy to implement and does not need a large amount of data to create. It is not impacted by missing values and the final model is easy to visualize with useful graphical representations. The results are easy to explain and is often analogous to human decision making compared to other models. The model does not work well with variables that are highly correlated and does not create a model with smooth boundaries. The greedy approach used will cause the models to have higher variance but helps to minimize what can become a complex model with high training. Trees are unstable, meaning small changes to the input can have large effects on the structure of the tree. Due to the hierarchical nature for tree growth, errors near the base effect the rest of the tree.

Table 4.3. *Building a Regression Tree Algorithm* (James et al, 2013)

1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α.
3. Use K-fold cross-validation to choose α. That is, divide the training observations into K folds. For each $k = 1, \ldots, K$:
    a. Repeat Steps 1 and 2 on all but the kth fold of the training data.
    b. Evaluate the mean squared prediction error on the data in the left-out kth fold, as a function of α.
    Average the results for each value of α, and pick α to minimize the average error.
4. Return the subtree from Step 2 that corresponds to the chosen value of α.

Random Forest

Random forest is a bagging tree method for building a regression model. Bagging, short for bootstrap aggregation, is a method to reduce variance in prediction errors. Bagging trees, or any high variance, low bias technique, improves predictive performance when compared to single trees. A random component is introduced when generating bootstrap samples. The general idea is to average many noisy but approximately unbiased models will lower variance and trees are a good choice because they can capture complex interactions in the data. If the trees are sufficiently deep, bias should be relatively low. When bagging, the trees are not completely independent because every parameter is used for each split and the bias will be the same as individual trees (Friedman et al, 2001). Reducing correlation among predictors can be done by adding randomness into the tree construction process (Kuhn & Johnson, 2013).

Random forest is a significant modification of bagging which decorrelates the trees created and then averages the outcomes. This is done through random selection of the input variables during the tree growing process. When growing a tree on the bootstrapped dataset, before each split a selection of $m \leq p$ of the variables are randomly selected as candidates. The general rule of thumb is $m = \sqrt{p}$ for classification and $m = \frac{1}{3}p$ for regression problems (Kuhn & Johnson, 2013). After some number of trees B are grown $\{T(x; \Theta_b)\}_1^B$ the predictor for the random forest regression is

$$\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T(x; \Theta_b).$$

Tables 4.4 and 4.5 are two algorithms that represent how random forests are created. Random forests are very accurate for many datasets where no feature scaling is required. It can handle many observations efficiently and handle many variables without variable deletion. The

model decorrelates the data and has lower variance than single trees. Missing data is also not a problem for random forests.

Random forests are often very accurate but are a hard to interpret black box approach, since they cannot be easily visualized like decision trees. Datasets with high amounts of noise can often be overfit. For data that includes categorical variables with different levels, the model will be biased in favor of attributes with more levels. While it may be an efficient model, it can be computationally intensive based on the size of the dataset and number of trees being created.

Table 4.4. *Random Forest Algorithm 1* (Kuhn & Johnson, 2013)

1. Select the number of models to build, m
2. for $i = 1$ to m do
    a. Generate a bootstrap sample of the original data
    b. Train a tree model on this sample
       for each split do
        i. Randomly select k $(< P)$ of the original predictors
        ii. Select the best predictor among the k predictors and partition the data
    c. end
3. Use typical tree model stopping criteria to determine when a tree is complete (but do not prune)

Table 4.5. *Random Forest Algorithm 2* (Friedman et al, 2001)

1. For $b = 1$ to B:
    a. Draw a bootstrap sample Z^* of size N for the training data.
    b. Grow a random-forest tree T_0 to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
        i. Select m variables at random from the p variables.
        ii. Pick the best variable/split-point among the m.
        iii. Split the node into two daughter nodes.
2. Output the ensemble of trees $\{T_b\}_1^B$.
3. To make a prediction at a new point x:
   Regression: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Gradient Boosting Machines (GBM)

Boosting models were originally created for classification problems and were later adapted for regression problems. Boosting assumes that there exists some base or weak learning algorithm which will create a weak classifier when given labeled training data. The goal is to take a weak learning algorithm and improve the performance. The assumption is that the base learner will produce a weak hypothesis that is at least better than a random guess. This is called the weak learning assumption that is the foundation of boosting (Schapire & Freund, 2013). These weak classifiers are combined, or boosted, to create an ensemble classifier. This ensemble classifier should have improved generalized error rate and can outperform stronger learners (Schapire, 1990). If the base learner is repeated on the same data, the results would not end up being interesting or different through iterations. This means the data being fed to the algorithm must be manipulated.

The initial success of the model, especially AdaBoost, led to many advances which eventually led to connecting the AdaBoost algorithm to statistical concepts such as loss functions, additive modeling, and logistic regression (Freidman et al, 2000). It was shown that boosting can be interpreted as a forward stepwise additive model that minimizes exponential loss and enabled the method to expand the approach to regression problems (Kuhn & Johnson, 2013).

The boosting approach learns slowly to prevent overfitting, as can happen with large decision trees, and simple classifiers have been shown to perform better (Freund & Schapire, 1996). The current tree is fit using the residuals from the previous model as the response instead of the outcome Y. The decision tree is included into the fitted function and the residuals are updated to repeat this process. Trees can be kept small to slow the learning process as well as the shrinkage parameter λ to limit overfitting (Friedman et al, 2000). While it is generally accepted

that a simpler classifier should limit generalized error rate, boosting methods can perform equally well as the combined classifier becomes more complex (Bartlett et all, 1998).

Gradient boosting machines is a highly adaptable algorithm for both classification and regression problems. The basic principles for the approach, seen in Table 4.7, is that given a loss function (e.g. squared error) and a weak learner (e.g. regression trees), the algorithm will find an additive model to minimize the loss function. A general guess, such as the mean, is selected to begin and the gradient (e.g. residual) is calculated and then a model is created to fit the residuals while minimizing the loss function. This model is added to the previous model and the steps are repeated until a set number of iterations have occurred. Because GBM is tasked with finding the optimal fit for each stage, it can lead to overfitting, so a greedy algorithm is applied to limit it. The greedy algorithm may not find the global optimal model, but this can be countered with regularization, or shrinkage.

Table 4.6. *Boosting Algorithm* (James et al, 2013)

1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set.
2. For $b = 1,2, \dots, B$, repeat:
    a. Fit a tree $\hat{f}^b$ with d splits ($d + 1$ terminal nodes) to the training data (X, r).
    b. Update $\hat{f}$ by adding in a shrinken version of the new tree:
    $$\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x).$$
    c. Update the residuals
    $$r_i \leftarrow r_i - \lambda \hat{f}^b(x_i).$$
3. Output the boosted model,
$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x).$$

Table 4.7. *Simple Gradient Boosting for Regression Algorithm* (Kuhn & Johnson, 2013)

1. Select tree depth D, and number of iterations, K
2. Compute the average response, $\bar{y}$, and use this as the initial predicted value for each sample
3. For $k = 1$ to K dp
    a. Compute the residual, the difference between the observed value and the current predicted value, for each sample
    b. Fit a regression tree of depth, D, using the residuals as the response
    c. Predict each sample using the regression tree fit in the previous step
    d. Update predicted value of each sample by adding the previous iteration's predicted value to the predicted value generated in the previous step
4. End

Steepest descent is used in gradient boosting machines. This chooses $h_m = -\rho_m g_m$ where ρ_m is a scalar and $g_m \in \mathbb{R}^N$. This is the gradient of $L(f)$ evaluated at $f = f_{m-1}$ with the components of the gradient

$$g_{im} = -\left[\frac{\partial L\big(y_i, f(x_i)\big)}{\partial f(x_i)}\right]_{f(x_i)=f_{m-1}(x_i)}.$$

This is used to find the step length ρ_m

$$\rho_m = argmin_\rho L(f_{m-1} - \rho g_m)$$

and the solution is updated

$$f_m = f_{m-1} - \rho_m g_m.$$

The negative gradient $-g_m$ defines the steepest descent, called the line search along the direction (Friedman, 2001). This is repeated for each iteration. This is a greedy method because $-g_m$ is the local direction of steepest descent. Given some given loss function $L(y, f)$ and a base learner h_m, the solution may be difficult to determine. To deal with this, we choose a new function that is the most parallel to the gradient with the observed data. This allows for the replacement of a

potential hard optimization problem with a classic least squares minimization (Natekin & Knoll, 2013; Friedman, 2002). Table 4.8 contains the algorithm for gradient boosting. Some common gradients can be found in Tables 4.9 and 4.10. To specify an arbitrary loss function, a function to calculate the corresponding negative gradient is required.

Table 4.8. *Gradient Boosting Algorithm* (Friedman et al, 2001)

1. Initialize $f_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$.
2. For $m = 1$ to M:
    a. For $i = 1,2, \dots, N$ compute
    $$r_{im} = -\left[\frac{\partial L\big(y_i, f(x_i)\big)}{\partial f(x_i)}\right]_{f=f_{m-1}}.$$
    b. Fit a regression tree to the targets r_{im} giving terminal regions $R_{jm}, j = 1,2, \dots, J_m$.
    c. For $j = 1,2, \dots, J_m$ compute
    $$\gamma_{jm} = argmin_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$
    d. Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I\big(x \in R_{jm}\big)$.
3. Output $\hat{f}(x) = f_M(x)$.

Gradient boosting machines are a powerful method for solving real world problems and can effectively work with complex non-linear data. They are flexible and can be adapted to many situations based on the chosen loss function and base learner. With this robust capability, the method is generally memory intensive because the model depends on the number of boosting iterations used to learn and build the model. Finding the optimal shrinkage parameter and working with a large number of parameters can make the model computationally difficult. This leads to a slower evaluation speed making online learning difficult without tradeoffs in accuracy. Because of the sequential nature of boosting, parallelization is not possible other than evaluation of already learned models. The computational difficulties are the tradeoff for such a powerful model that is highly applicable.

Table 4.9. *Gradients* (Natekin & Knoll, 2013)

1. Continuous response, $y \in \mathbb{R}$:
    a. Gaussian L_2 loss function
    b. Laplace L_1 loss function
    c. Huber loss function, δ specified
    d. Quantile loss function, α specified
2. Categorical response $y \in \{0,1\}$:
    a. Binomial loss function
    b. AdaBoost loss function
3. Other families of response variable:
    a. Loss functions for survival models
    b. Loss functions counts data
    c. Custom loss functions

Table 4.10. *Loss Functions*

Setting	Loss Function	$\partial L(y_i, f(x_i))/\partial f(x_i)$
Regression	$\frac{1}{2}[y_i - f(x_i)]^2$	$y_i - f(x_i)$
Regression	$\lvert y_i - f(x_i)\rvert$	$sign[y_i - f(x_i)]$
Regression	Huber	$y_i - f(x_i)\ for\ \lvert y_i - f(x_i)\rvert \le \delta_m$ $\delta_m sign[y_i - f(x_i)]\ for\ \lvert y_i - f(x_i)\rvert > \delta_m$ $where\ \delta_m = \alpha th - quantile\{\lvert y_i - f(x_i)\rvert\}$
Classification	Deviance	kth component: $I(y_i = \mathcal{G}_k) - p_k(x_i)$

Stochastic Gradient Boosting Machines

Stochastic gradient boosting is a modification of gradient boosting that incorporates randomness as part of the procedure. During each iteration a subset is drawn at random without replacement instead of using the entire dataset. This subset is used to fit the base learner and computes the current iteration. As before, $\{y_i, x_i\}_1^N$ represents the entire training data sample and we will consider a random permutation of the integers $\{1, \dots, N\}$ written as $\{\pi(i)\}_1^N$. Based on this, there will be a subsample of size $\widetilde{N} < N$ is written as $\left\{y_{\pi(i)}, x_{\pi(i)}\right\}_1^{\widetilde{N}}$. If $\widetilde{N} = N$, then no randomness will be included, and it will be the same as gradient boosting. As the fraction of the data that is included decreases, the randomness will increase, but limits the amount of available data to the learner at each iteration. This will increase the variance associated with the estimates

by base learners. Including the randomly selected subsets in the gradient boosting algorithm, Table 4.11 is the algorithm adapted for stochastic gradient boosting.

Table 4.11. *Stochastic Gradient Boosting Algorithm* (Friedman, 2002)

1. $F_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$
2. For $m = 1$ to M do:
    a. $\{\pi(i)\}_1^N = rand_{perm}\{i\}_1^N$
    b. $\tilde{y}_{\pi(i)m} == -\left[\frac{\partial L\left(y_{\pi(i)}, f\left(x_{\pi(i)}\right)\right)}{\partial f\left(x_{\pi(i)}\right)}\right]_{f(x)=f_{m-1}(x)}, i = 1, \tilde{N}$
    c. $\{R_{lm}\}_1^L = L - terminal\ node\ tree\ \left(\{\tilde{y}_{\pi(i)m}, x_{\pi(i)}\}_1^{\tilde{N}}\right)$
    d. $\gamma_{lm} = argmin_\gamma \sum_{x_{\pi(i)} \in R_{lm}} L\left(y_{\pi(i)}, f_{m-1}\left(x_{\pi(i)}\right) + \gamma\right)$
    e. $f_m(x) = f_{m-1}(x) + \upsilon \cdot \gamma_{lm} 1(x \in R_{lm})$
3. endFor

Choosing a Model

Various models have been introduced here and it is important to consider which model may be the best choice. To do this, a comparison of assumptions, strengths, and weaknesses should be considered. The ILIAD data has moderate multicollinearity, which the shrinkage methods deal well with. Of the shrinkage methods, Ridge regression does not perform variable deletion. The data contains a limited number of lexical characteristics, so variable deletion likely won't have a large impact. All of the shrinkage methods have many of the assumptions of linear regressions, which we have demonstrated the ILIAD data does not follow. Because of this, the shrinkage methods are unlikely to perform as well as the other advanced methods.

Support vector regressions are very adaptable based on the kernel selection. The kernel chosen will determine the assumptions for the model and some prior expertise may help with its selection. Support vectors assume the data are independent and identically distributed, which not true for the ILIAD data so may impact the final model. Regression trees are similarly adaptable but the greedy nature may not lead to the best model based on the multicollinearity of the data.

Random forests, gradient boosting machines, and stochastic gradient boosting machines rely of very few assumptions. Like support vectors, gradient boosting is reliant on the choice of gradient or descent method. Some expertise aids in this decision and the choice of gradient can input some assumptions about the data. Random forest has no formal distributional assumption and decorrelates the data, make it a very strong model. Stochastic gradient boosting machines takes the strengths of gradient boosting and augments them by including randomness. This randomness decorrelates the variables, much like random forest. While certain models may be significantly stronger based on the underlying theory, the most important factor in choosing a model is what it demonstrates for researchers. Sometimes the best model is not the most accurate model, if no understanding can be gained from the results.

In this chapter, we have contributed to word learning research by introducing alternative modeling techniques to analyze educational data. This survey of models included a brief overview of each method to explain the theory and limitations, as well comparing their strengths for word learning research. The comparisons were made using the limitations found exploring the ILIAD data as an example dataset.

CHAPTER FIVE:

CHOOSING MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS): A MODEL COMPARISON

Multivariate Adaptive Regression Splines (MARS)

In the previous chapter we introduced advanced statistical learning and machine learning models that are more robust and can improve the understanding of the ILIAD word learning data. Based on the ILIAD data and its underlying conditions (multicollinearity, skew, heteroscedasticity, non-normal parameters), another method may be a better alternative by balancing being robust and adaptable while also being interpretable and easier to implement.

Multivariate adaptive regression splines (MARS) does not rely on the same assumptions as linear models like multiple and stepwise regressions, such as a linear relationship or homoscedasticity, and provides nuanced information about the relationships modeled. MARS is a general additive model (GAM) that was first introduced by Friedman (1991). It uses recursive partitioning of the data with hinges or splines, motivated by classification and regression trees (CART), that can capture higher order interactions with more power and flexibility (Friedman & Roosen, 1995). MARS handles nonlinearity well, whereas multivariate linear regression, stepwise regressions, and shrinkage methods assume linearity. Through adaptive and automatic variable selection and iterative partitioning, MARS does particularly well when dealing with high dimensional data by partitioning the data into smaller subsections. This can be viewed as a geometrical procedure or a generalization of stepwise linear regression. It does this by additively using piecewise linear basis functions. MARS was considered for modelling the ILIAD data

because it was believed the hinged data would more accurately describe word learning than linear models and its adaptable nature for the complexities of data from young children.

The recursive partitioning for regression model is made up of $\{R_m\}_1^M$ disjoint subregions representing a partition of D, the domain of the independent variables. The goal is to split the data into subregions and estimate the parameters associated with the separate functions for each individual subregion. The starting region is the entire domain D, which is then recursively split among subregions. These subregions allow us to more precisely measure word learning by determining thresholds where significant changes occur, such as the age of acquisition where word learning begins to decrease for students in first grade. MARS is an expansion of basis functions and the associated coefficients $\{a_m\}_1^M$. These basis functions, B_m, take the form

$$B_m(x) = I[x \in R_m],$$

where the indicator function *I* has a value of one if the argument is true and zero otherwise. This is a generalization of stepwise regression where the step function takes the form

$$\hat{f}(x) = \sum_{m=1}^{M} a_m B_m(x).$$

MARS differentiates itself from stepwise regression at this stage and adapts to nonparametric data by dividing subregions using hinges. These subregions are disjoint, so only one basis function may be nonzero for any point x. Subregions are optimally split into two linear functions called reflected pairs, R_l and R_r. These piecewise linear basis functions are split at the hinge t and take the form $(x-t)_+$ and $(t-x)_+$, where + is the positive part.

$$(x-t)_+ = \begin{cases} x-t, & if\ x > t \\ 0, & otherwise \end{cases} \text{ and } (t-x)_+ = \begin{cases} t-x, & if\ x < t \\ 0, & otherwise \end{cases}.$$

Optimal hinge placement is achieved by minimizing root mean square error (RMSE), where

$$RMSE(\hat{\theta}) = \sqrt{E(\hat{\theta} - \theta)^2},$$

where θ represents the parameter of interest and $\hat{\theta}$ is the associated estimate. A collection of the basis functions of the form

$$C = \left\{(X_j - t)_+, (t - X_j)_+\right\} \text{ where } t \in \{x_{1j}, x_{2j}, \dots, x_{Nj}\} \text{ and } j = 1,2, \dots$$

is created with the input vector of interest X_j and a given spline t that cuts the data at a given observation x_{ij}. A spline or knot is created at each point x_{ij} and reflected pairs are formed at each observed value for that input to build the collection of possible basis functions. Functions from the set C and their products will be used instead of the original inputs where each input can appear at most in one product. With this in mind, the model has the form

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X),$$

where $h_m(X)$ is a function in C, or a product of such functions and X is the data made up of the input vectors X_j.

With the choice of $h_m(X)$, the coefficients β_m can be estimated by minimizing the residual sum-of-squares of linear regressions in each candidate subregion. The product of a function h_m and one of the reflected pairs in C is considered for a new basis function pair at each step. This term can be added to the model $\mathcal{M}$ and the term that produces the largest decrease in training error is chosen. The algorithmic steps for splitting the data and choosing the coefficients for each subregion are repeated until a predetermined maximum number of terms in the model $\mathcal{M}_{max}$.

To limit overfitting, backwards deletion is a procedure to remove the term that causes the smallest increase in residual squared error. This is done to produce an estimated best model $\hat{f}_\lambda$ of each number of terms λ, contingent on the total number of variables being considered for the model. Cross validation can be used to determine the optimal λ and saves computational time. The MARS procedure uses generalized cross validation defined as

$$GCV(\lambda) = \frac{\sum_{i=1}^{N}\left(y_i - \hat{f}_\lambda(x_i)\right)^2}{(1 - M(\lambda)/N)^2}$$

where $M(\lambda)$ is the effective number of parameters in the model. This accounts for both the number of parameters used in selecting the optimal positions of knots and the number of terms in the model (Hastie, Tibshirani, & Friedman, 2001).

The data being analyzed is split between first, second, and third graders. A categorical variable for grade was included to represent these groups. For categorical variables, another set of basis functions need to be used. As with the original formulation, a set of basis functions are derived by taking the tensor product over all of the variables of the univariate basis functions

$$I\left(x_j \in A_{\ell j}\right), \quad 1 \leq j \leq n.$$

An adaptive strategy must be applied that would consider all basis functions in the complete tensor product as candidate variables similarly to ordinal subsets, but the truncated power splines will be replaced by indicator functions over the categorical variable subsets (Friedman, 1991)

$$[+(x_v - t)]_+^q \leftarrow I(x_v \in A)$$
$$[-(x_v - t)]_+^q \leftarrow I(x_v \notin A)$$

One of the major strengths of MARS is that it does not begin with any distributional assumptions. It is a non-parametric regression model, so it does not rely on data from a normal distribution. It handles correlated variables relatively well and works well with a large number of

predictor variables. The iterative nature of MARS allows it to handle both linear and non-linear data. The additive nature of the model allows the effect of each variable X_j on Y can be examined while keeping all other variables fixed (James et al., 2013). Because of the nature of its hinges, MARS is robust to outliers because it can place a hinge at the outlier to limit the impact. When compared to other advanced modeling techniques, MARS is a fast, efficient algorithm given its complexity and can automatically detect interactions between variables.

While MARS has many strengths, some limitations do exist. It is slower to train than some simpler models because generalized linear models can be computationally expensive with high dimensional inputs (Murphy, 2012). While it handles correlated predictors well, this can make model interpretation difficult. If two variables are highly correlated, the first one encountered by the model will be chosen using a greedy method (Boehmke & Greenwell, 2019). MARS can be sensitive to multicollinearity, but tests have shown that it performs well when compared to many other statistical methods (Dormann et al., 2013). Overfitting can be a problem, though a backward algorithm can limit this. It does not do well with missing data. Another issue that can be a shortcoming is the selection of candidate basis functions. The default basis functions for the model are linear, while others are available (e.g., polynomial basis functions), they must be manually selected by the researcher. Basis functions and knot locations can be computed automatically with advanced methods, such as an empirical Bayes method (Sakamoto, 2007), and the determining knot locations can be optimized with hill climbing methods (HCM, Ju et al, 2021).

MARS has been used to examine academic achievement (Kilc Depren, 2018; Martis et al., 2015) and to determine factors that impact early childhood development (Kolyshkina et al., 2013). It has not been applied to word learning and word recognition research. The purpose of

this study is to examine the relations between different lexical characteristics and word learning using MARS. We compare the results produced by MARS with multivariate linear regression and stepwise regression to demonstrate the added benefit of using a more adaptable statistical model like MARS.

The steps for creating a regression using multivariate adaptive regression splines are split into three algorithms. The first, found in Table 5.1, is a recursive partitioning algorithm that sets up how the data will be partitioned and the optimal split position is made. The forward stepwise algorithm, in Table 5.2, shows how the regression model is made. After the initial regression, a backward algorithm, found in Table 5.3, removes partitions and fills the gaps in order to create a simpler model and limit overfitting. For the algorithm in Table 5.1, the $LOF(g)$ is a procedure that computes the lack-of-fit of a function $g(x)$ to the data, for example minimizing residual sum of squares in linear piecewise regression. The following algorithm outlines the recursive partitioning strategy. The algorithm starts with the first line setting the initial region to the entire domain. The next step is a loop that iterates a splitting procedure with M_{max} as the final number of basis functions. The next three loops perform an optimization procedure to select the basis function B_{m^*}, a predictor variable x_{v^*}, and a split point t^*. The lack-of-fit is minimized for the model with B_{m^*} replaced by its product with the step function $H[+(x_{v^*} - t^*)]$ and an additional term that is the product of B_{m^*} and the reflected step function $H[-(x_{v^*} - t^*)]$. This means it splits the corresponding region R_{m^*} on variable v^* at split point t^*. The minimization of $LOF(g)$ is a linear regression of the response on the current basis function set with respect to the expansion coefficients.

Table 5.1. *Recursive Partitioning Algorithm* (Friedman, 1991)

$B_1(x) \leftarrow 1$
$For\ M = 2\ to\ M_{max}\ do: lof^* \leftarrow \infty$
$For\ m = 1\ to\ M - 1\ do:$
$For\ v = 1\ to\ n\ do:$
$For\ t \in \{x_{vj} | B_m(x_j) > 0\}$
$g \leftarrow \sum_{i \neq m} a_i B_i(x) + a_m B_m(x) H[+(x_v - t)] + a_m B_m(x) H[-(x_v - t)]$
$lof \leftarrow min_{a_1,\ldots,a_M} LOF(g)$
$if\ lof < lof^*, then\ lof^* \leftarrow lof; m^* \leftarrow m; v^* \leftarrow v; t^* \leftarrow t\ end\ if$
$end\ for$
$end\ for$
$B_M(x) \leftarrow B_{m^*}(x) H[-(x_{v^*} - t^*)]$
$B_{m^*}(x) \leftarrow B_{m^*}(x) H[+(x_{v^*} - t^*)]$
$end\ for$
$end\ algorithm$

Table 5.2. *MARS Forward Stepwise Algorithm* (Friendman, 1991)

$B_1(x) \leftarrow 1; M \leftarrow 2$
$Loop\ until\ M > M_{max}: lof^* \leftarrow \infty$
$For\ m = 1\ to\ M - 1\ do:$
$For\ v \notin \{v(k,m) | 1 \leq k \leq K_m\}$
$For\ t \in \{x_{vj} | B_m(x_j) > 0\}$
$g \leftarrow \sum_{i=1}^{M-1} a_i B_i(x) + a_M B_m(x)[+(x_v - t)]_+ + a_{M+1} B_m(x)[-(x_v - t)]_+$
$lof \leftarrow min_{a_1,\ldots,a_{M+1}} LOF(g)$
$if\ lof < lof^*, then\ lof^* \leftarrow lof; m^* \leftarrow m; v^* \leftarrow v; t^* \leftarrow t\ end\ if$
$end\ for$
$end\ for$
$end\ for$
$B_M(x) \leftarrow B_{m^*}(x)[+(x_{v^*} - t^*)]_+$
$B_{M+1}(x) \leftarrow B_{m^*}(x)[-(x_{v^*} - t^*)]_+$
$M \leftarrow M + 2$
$end\ loop$
$end\ algorithm$

After the forward stepwise algorithm found in Table 5.2 is initiated, a backward stepwise algorithm is necessary. This is to limit overfitting in the model which would likely occur if only forward stepwise was completed. This backward stepwise algorithm can be found in Table 5.3.

Table 5.3. *MARS Backwards Stepwise Algorithm* (Friedman, 1991)

$J^* = \{1,2,\ldots,M_{max}\}; K^* \leftarrow J^*$
$lof^* \leftarrow min_{\{a_j|j\in J^*\}} LOF\left(\sum_{j\in J^*} a_j B_j(x)\right)$
$For\ M = M_{max}\ to\ 2\ do: b \leftarrow \infty; L \leftarrow K^*$
$For\ m = 2\ to\ M\ do: K \leftarrow L - \{m\}$
$lof \leftarrow \min_{\{a_k|k\in K\}} LOF(\sum_{k\in K} a_k B_k(x))$
$if\ lof < b, then\ b \leftarrow lof; K^* \leftarrow K\ end\ if$
$if\ lof < lof^*, then\ lof^* \leftarrow lof; J^* \leftarrow K\ end\ if$
$end\ for$
$end\ for$
$end\ algorithm$

After the algorithm from Table 5.2 is used, the initial model is comprised of the basis function set $J*$. During each iteration of the outer For loop in the algorithm from Table 5.3, one basis function is deleted from the model. The inner For loop chooses which basis function is selected for deletion. This selection is the basis function whose removal improves the fit most or degrades it the least. During these For loops, the basis function $B_1(x) = 1$ is never removed. The algorithm constructs a sequence of $M_{max} - 1$ models, with each step in the sequence having one less basis function. The best model in the sequence is returned when the algorithm is complete.

Model Comparison Results

The ILIAD dataset was used to model the relation between lexical characteristics and students' word learning. This data consisted of 350 students' decontextualized word learning and expressive labeling outcomes for 377 words and their respective lexical characteristics across grade levels. Each word was characterized by its lexical characteristics for individual word frequency, age of acquisition, level of concreteness, phonological neighborhood density, and phonotactic probability. Preliminary data analyses determined the data did not follow a normal distribution, the data was right skewed, and moderate multicollinearity existed among the independent variables.

Each model examined the influence of lexical characteristics on decontextualized word learning and expressive labeling for each grade level (first, second, and third grades) and a full model that included all grade levels and a categorical variable for grade level. The RStudio environment was used to create the models in R. Base R was used for multivariate linear regression, and the MASS package was used to complete the stepwise regression analysis (Venables and Ripley, 2002). For the shrinkage methods, ridge regression and LASSO were completed using the glmnet package (Friedman et al, 2010) and partial least squares and principal component regression used the pls package (Wehrens & Mevik, 2007). Support vector regression was done using the e1071 package (Meyer et al, 2019). MARS is a trademarked name, so the package used in R is called Earth (enhanced adaptive regression through hinges) for the research (Milborrow, 2011). For tree-based methods, regression trees were created using the rpart package (Therneau & Atkinson, 2019), random forest using the randomForest package (Liaw & Wiener, 2002), and gradient boosting machines using the gbm package (Greenwell et al, 2020).

To determine how well MARS regresses the data and why it is a great choice for researchers, the ILIAD data needed to be modeled by each technique and compared. Comparisons between models were completed using measures of accuracy including R^2, mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE). These metrics were chosen to aid in comparison with the MARS model by mimicking prior analysis in educational methods. Any other fit metrics that the models rely on were also included for the sake of thoroughness, such as ridge traces, variable importance, and graphs of accuracy as more trees are included in an ensemble method.

Each model was compared using R^2, mean absolute error (MAE), mean square error (MSE), and root mean square error (RMSE). These fit metrics were chosen for ease of comparison for model performance based on the variety of models being considered. While creating each model, other fit metrics were used to create the strongest models. These include model specific steps and metrics such as ridge tracing, cross validation, or parameter tuning. Each individual model discusses the special metrics but for uniformity comparisons were based on R^2 and error metrics for fit.

R^2, the coefficient of multiple determination, represents the proportion of variance in the dependent variables, $y_i's$, which is accounted for by the independent variables, $x_1, \dots, x_k$ (Johnson & Wichern, 2014). This means it acts as a goodness of fit indicator where values range from 0 to 1 with higher values indicating a larger effect. If the R^2 values are close to 1, the model can be considered a good fit (Rencher & Schaalje, 2008). R^2 is the proportion of the sum of squares of deviations of the output values, which represent the linear relationship between independent and dependent variables (Ramachandran & Tsokos, 2020). R^2 is not designed to work with nonlinear data meaning it may not be reliable as the sole metric for model fit but it is a commonly reported metric that can still serve as a useful summary statistic for measuring model accuracy (Kvålseth, 1985). The coefficient of determination can be calculated with

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = \frac{\hat{\beta}_1' X_c' X_c \hat{\beta}_1}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = \frac{SSR}{SST}.$$

Mean absolute error, mean square error, and root mean square error are all measures to compare the error between predicted results and the true outcomes. Because they are error measures, values that are closer to zero indicate a stronger model (Tamhane & Dunlop, 2000). Mean absolute error can be found with

$$MAE = \frac{\sum_{i=1}^{n}|\varepsilon|}{n} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n},$$

mean square error with

$$MSE = \frac{\sum_{i=1}^{n}(\varepsilon^2)}{n} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n},$$

and root mean square error with

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\varepsilon^2)}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}},$$

where ε is the error, which represents the difference from the observations from the predicted values. All three of these methods measure the error, or distance of the predicted values from the true values, by measuring the distance and making all outcomes positive through the absolute value or squaring the values. These are then summed over the number of observations. It has been argued that MAE is a better alternative to RMSE because it is "a more natural" measure of average error and is unambiguous (Willmott & Matsuura, 2005).

Multivariate linear regression and stepwise regression were included because they are often used by educational researchers and act as a baseline. Ridge regression, LASSO, PLS, and PCR were included to determine to impact multicollinearity had on regular regression methods. Support vector regression was used to model the data because it has similar capability to model nonlinear data to MARS but may be harder to interpret the resulting model (Drucker et al, 1997). Regression trees were included because they are building blocks of ensemble tree-based methods and are readily interpretable. The ensemble tree-based methods were included because they do not rely on data having a normal distribution, but the results are difficult to interpret and are often seen as "black box" models.

Model results for word learning measured by the decontextualized and expressive tasks are listed in Tables 5.4, and 5.5, respectively. The tables have subsections for the full model, first grade, second grade, and third grade. Each model is compared using fit metrics to determine how accurately the data was explained.

Generally, the shrinkage methods performed as well as multivariate linear regression and stepwise regression according to the error measurements and R^2, for both decontextualized learning and expressive labeling. Ridge regression, LASSO, and PCR had error measures comparable to multivariate linear regression but accounted for less of the variance $\left(R^2_{ridge} = .14, R^2_{LASSO} = .26, R^2_{PCR} = .20\right)$. This was true for the third-grade model with expressive labeling outputs $\left(R^2_{ridge} = .06, R^2_{LASSO} = .22, R^2_{PCR} = .22\right)$. Partial least squares maintained an R^2 value close to multivariate linear regression.

Support vector regression and regression trees performed comparable to MARS. This is true across all grade levels for both decontextualized learning and expressive tasks. By every metric considered, SVR is slightly better than regression trees. MARS explains more of the variance in most models compared to SVR, but SVR often have slightly better error measures. One exception was the second-grade model for expressive labeling, where SVR had a higher R^2 than any other method. This was expected because prior experiments have compared support vector regression to MARS and found that MARS has a higher modeling error than SVR more often than not, but prediction error was similar (Drucker et al, 1997). They determined that the test data may have been too simple, so it seemed appropriate to make a similar comparison with data that is not as streamlined. Support vector regressions are more restrictive for word learning analysis because it assumes that the data is independent and identically distributed (i.i.d., Awad & Khanna, 2015).

Gradient boosting machines and stochastic gradient boosting machines explained more of the variance in the dependent variable than the other models. According to fit metrics, the gradient boosting methods fit the data better compared to other methods with the exception of random forest. Random forest had a similar R^2 to regular regression but significantly better fit.

For first-grade models, stochastic gradient boosting had the highest R^2 for decontextualized learning and gradient boosting had the highest R^2 for expressive labeling. Random forest had the best fit metrics for both models and had an R^2 similar to MARS. For the second-grade models, the results for decontextualized were similar to the first-grade model for model fits. For expressive, support vector regression had a higher R^2 other models, including gradient boosting machines. Random forest continued to have the strongest fit error metrics and the shrinkage methods did as well as multivariate regression. Stochastic gradient boosting machines performed the best by every metric for the third-grade model using decontextualized word learning. For expressive, gradient boosting had the highest R^2 and random forest performed the best according to MAE, MSE, and RMSE. Overall, stochastic gradient boosting, gradient boosting machines, and random forest performed better than the other methods considered.

The shrinkage models were the weakest across all grade levels for both decontextualized and expressive learning. Ridge regression, LASSO, partial least squares, and principal component regression consistently fit the data as well as multivariate linear regression for both R^2 and the error measures. This does not mean these models are bad. The shrinkage methods' strength is the ability to handle multicollinearity. Having similar results to multivariate linear regression shows that while the data had some multicollinearity, the impact from it is not overly great. It is interesting to note that the shrinkage methods performed comparably to each other except for third grade. The third-grade learning results were the most inconsistent, so the

variability explained by each model deviated from each other. Partial least squares did the best under those circumstances.

Regression trees consistently outperformed shrinkage models across each grade for both decontextualized learning and expressive task labeling. The regression trees were generally close in fit the MARS, with MARS doing slightly better overall. Support vector regression explained the dataset as well as regression trees overall but explained the most variability for the expressive second and third grade models.

Random forest had the best error metrics for every grade level for both decontextualized learning and expressive labeling tasks, except for the third-grade decontextualized model. For the third-grade models, random forest explained less variability than regression trees and support vector regression and far less than gradient boosting machines for decontextualized learning which had an R^2 of .91. For error metrics is performed the best overall for the third-grade expressive model but slightly less than gradient boosting machines for decontextualized learning. Gradient boosting machines explained the most variability for all decontextualized models and the full and first grade models for expressive labelling. For the second-grade expressive model support vector machines had a higher R^2 and for the third-grade model gradient boosting was one of the lowest performing models.

To explain some of the variability in the models' performances, it is important to consider the differences between learning measures. Both assessed children's word knowledge, but in different ways. The decontextualized measure assessed their ability to define vocabulary words without any other context provided, and the expressive measure assessed children's ability

Table 5.4. *Model Comparison for Decontextualized Word Learning*

	MARS	MR	SR	Ridge	LASSO	PLS	PCR	SVR	CART	RF	GBM	SGBM
Full Model												
R^2	.43	.31	.31	.30	.30	.31	.30	.39	.41	.31	.42	**.52**
MAE	.15	.17	.17	.14	.14	.17	.17	.13	.15	**.07**	.15	.13
MSE	.04	.05	.03	.04	.04	.05	.05	.04	.04	**.01**	.04	.03
RMSE	.20	.22	.22	.19	.19	.22	.22	.20	.20	**.10**	.20	.18
First Grade												
R^2	.84	.69	.69	.69	.68	.69	.63	.81	.77	.81	.88	**.90**
MAE	.09	.13	.13	.13	.13	.13	.13	.08	.10	**.04**	.07	.07
MSE	.01	.03	.03	.03	.03	.03	.03	.02	.02	**.00**	.01	.01
RMSE	.12	.16	.16	.16	.16	.16	.17	.13	.13	**.06**	.10	.09
Second Grade												
R^2	.72	.58	.57	.53	.57	.58	.58	.69	.66	.60	.73	**.84**
MAE	.12	.15	.15	.16	.15	.15	.15	.11	.12	**.06**	.12	.09
MSE	.02	.03	.03	.04	.03	.03	.03	.02	.03	**.01**	.02	.01
RMSE	.15	.18	.19	.19	.19	.18	.18	.16	.17	**.08**	.15	.11
Third Grade												
R^2	.42	.33	.31	.14	.26	.32	.20	.41	.35	.28	**.94**	.91
MAE	.08	.09	.09	.10	.10	.09	.10	.08	.09	.04	**.03**	**.03**
MSE	.01	.02	.01	.02	.02	.01	.02	.01	.01	.00	**.00**	**.00**
RMSE	.11	.12	.12	.14	.13	.12	.13	.11	.12	.06	**.04**	**.04**

Note. MARS= Multivariate Adaptive Regression Splines, MR= multivariate linear regression, SR= Stepwise Regression, LASSO= Least Absolute Shrinkage & Selection Operator, PLS= Partial Least Squares, PCR= Principal Component Regression, SVR= Support Vector Regression, CART= Classification & Regression Tree, RF= Random Forest, GBM= Gradient Boosting Machines, SGBM= Stochastic Gradient Boosting Machines, MAE= Mean Absolute Error, MSE= Mean Square Error, RMSE= Root Mean Square Error.

Table 5.5. *Model Comparison for Expressive Word Learning*

	MARS	MR	SR	Ridge	LASSO	PLS	PCR	SVR	CART	RF	GBM	SGBM
Full Model												
R^2	.42	.30	.30	.31	.31	.30	.30	.41	.38	.30	.44	**.53**
MAE	.12	.14	.17	.17	.17	.14	.14	.12	.13	**.06**	.12	.11
MSE	.03	.04	.05	.05	.05	.04	.04	.03	.03	**.01**	.03	.02
RMSE	.17	.19	.22	.22	.22	.19	.19	.17	.18	**.09**	.17	.15
First Grade												
R^2	.86	.68	.68	.68	.68	.68	.67	.83	.86	.83	**.91**	.90
MAE	.07	.11	.13	.11	.11	.11	.11	.07	.07	**.03**	.05	.06
MSE	.01	.02	.03	.02	.02	.02	.12	.01	.01	**.00**	.01	.01
RMSE	.09	.14	.17	.14	.14	.14	.14	.10	.09	**.05**	.07	.08
Second Grade												
R^2	.64	.50	.49	.48	.48	.50	.50	**.68**	.59	.48	.58	.62
MAE	.11	.14	.16	.14	.14	.14	.14	.10	.12	**.06**	.12	.11
MSE	.02	.03	.04	.03	.03	.03	.03	.02	.02	**.01**	.02	.02
RMSE	.14	.17	.19	.17	.17	.13	.17	.13	.15	**.08**	.15	.15
Third Grade												
R^2	.40	.33	.32	.06	.22	.30	.22	.41	.38	.29	**.70**	.18
MAE	.08	.08	.11	.10	.09	.09	.09	.07	.08	**.04**	.05	.09
MSE	.01	.01	.02	.02	.12	.01	.01	.01	.01	**.00**	.01	.01
RMSE	.10	.11	.13	.13	.12	.11	.12	.10	.10	**.05**	.07	.12

Note. MARS= Multivariate Adaptive Regression Splines, MR= multivariate linear regression, SR= Stepwise Regression, LASSO= Least Absolute Shrinkage & Selection Operator, PLS= Partial Least Squares, PCR= Principal Component Regression, SVR= Support Vector Regression, CART= Classification & Regression Tree, RF= Random Forest, GBM= Gradient Boosting Machines, SGBM= Stochastic Gradient Boosting Machines, MAE= Mean Absolute Error, MSE= Mean Square Error, RMSE= Root Mean Square Error.

to use the vocabulary word to label a picture (Goldstein et al., 2017). While each measure taps into a different level of knowledge, decontextualized learning results may be more reliable for the models because it requires children to use higher-order processing skills to demonstrate comprehension of words (McKeown & Beck, 2014). Looking at the data itself, the third-grade results were the most inconsistent and were the biggest challenge for the models. Gradient boosting of regression trees is especially appropriate for mining less than clean data (Friedman, 2001) but ensemble methods often require more data to perform well, though bootstrapping can help. The third-grade dataset contains outcomes for 108 words, which may be too small to create robust ensemble models.

Based on the model comparisons, support vector regression, regression trees, random forest, and gradient boosting machines performed well overall. Gradient boosting machines explained the variability in the response well and random forest consistently had minimal error. While random forest and gradient boosting outdid other models most of the time, when they failed, they dropped well below more consistent models. This may be due to the requirement for more data to maintain reliability when using ensemble methods. Support vectors and regression trees performed well overall, and the results are more interpretable than ensemble methods. Comparing the results with multivariate adaptive regression splines, MARS slightly outperformed support vector regression and regression trees, was more consistent than the ensemble methods, and is more interpretable than the tree-based methods.

Computation Time

The computational complexity of algorithms is an important factor when considering which to use. Alan Turing investigated the computability of sequences and showed that a set of sequences can be partitioned into computable and non-computable sequences (Turning, 1937).

The computable sequences, which may be easy to compute or very complex, are classified based on computational complexity (Hartmanis & Stearns, 1965). As the scale of data grows, the chosen model can become increasingly time consuming. There are different formulations to compare the computational complexity of algorithms, such as big omicron (or big O), big omega, and big theta. Big O, the most commonly used, is an upper bound on the computational complexity based on the required steps. That is, $|f|$ is bounded above by g (a constant factor) asymptotically. $f(n) = \mathrm{O}\big(g(n)\big)$ where

$$\lim\ sup_{n\to\infty} \frac{|f(n)|}{g(n)} < \infty.$$

Big omega is a lower bound on the computational complexity $f(n) = \Omega\big(g(n)\big)$ where

$$\lim\ inf_{n\to\infty} \frac{f(n)}{g(n)} > 0.$$

Big theta is bounded above and below asymptotically, where $f(n) = \Theta\big(g(n)\big)$ falls between $\mathrm{O}\big(g(n)\big)$ and $\Omega\big(g(n)\big)$ (Knuth, 1976).

Many problems that machine learning and artificial intelligence systems are applied to are extremely complex and poorly understood, to the point of being beyond mathematical formalization (Kearns, 1990). This is because the formal definitions rely on specific operations while many machine learning algorithms have an undefined number of operations until some convergence criteria is met. Because of this, an operation like support vector regression will fall somewhere between $\mathrm{O}(n^2)$ and $\mathrm{O}(n^3)$ based on the kernel chosen (Bottou & Lin, 2007). This may not be a problem for smaller datasets but support vector machines scale rather badly due to the quadratic optimization algorithm and kernel transformation (Meyer & Wien, 2021).

Another way to consider the problem is through computational time. While computational complexity considers the set operations and their theoretical complexity,

computational time is based on empirical results that take into account efficiency of the algorithm and its implementation. The computational times for the compared methods can be found in Table 5.6. The computational time was based on the time to train each model and time to make a prediction based on the model.

For the training times, each model creation was replicated 20 times and the total time in seconds was documented and relative times were computed based on the fastest algorithm. For total elapsed time, multivariate linear regression took .03 seconds and was the fastest training model. Setting that to 1, the time of every other method was used to find the relative speed of the algorithm. The methods with the lowest computation time after multivariate linear regression were support vector regression ($t_{rel} = 1.33$), regression trees ($t_{rel} = 4.67$), MARS ($t_{rel} = 5.67$), and principal component regression ($t_{rel} = 6.33$). Support vector regression is very dependent on the chosen kernel and Gaussian, polynomial, and sigmoid are not very computationally complex. MARS was one of the faster models because it is a very efficient algorithm given the complexity of its operations.

Prediction times were based on 1,000 replications and the total time and relative time to the fastest prediction were similarly documented. The fastest predictions were ridge ($t_{rel} = 1$), LASSO ($t_{rel} = 1.03$), multivariate linear regression and support vector regression ($t_{rel} = 2.52$), and gradient boosting machines ($t_{rel} = 3.26$). MARS was the fourth slowest method for prediction, only being faster than stochastic gradient boosting machines ($t_{rel} = 6.58$), stepwise regression ($t_{rel} = 23.65$), and random forest ($t_{rel} = 61.74$). Prediction times are very small compared to training time, so they are not as consequential for model selection.

Table 5.6. *Computational Time*

Algorithm	Training Time (Seconds)	Training Time (Relative)	Prediction Time (Seconds)	Prediction Time (Relative)
MR	.03	1	.78	2.52
SR	30.05	1001.67	7.33	23.65
MARS	.17	5.67	1.63	5.26
Ridge	1.8	60	.31	1
LASSO	1.53	51	.32	1.03
PLS	.2	6.67	1.06	3.42
PCR	.19	6.33	1.08	3.48
SVR	.04	1.33	.78	2.52
CART	.14	4.67	1.25	4.03
RF	8.89	299.33	19.14	61.74
GBM	343.04	11,434.67	1.01	3.26
SGBM	1,215.03	40,501	2.04	6.58

Note. MR= multivariate linear regression, SR= Stepwise Regression, MARS= Multivariate Adaptive Regression Splines, LASSO= Least Absolute Shrinkage & Selection Operator, PLS= Partial Least Squares, PCR= Principal Component Regression, SVR= Support Vector Regression, CART= Classification & Regression Tree, RF= Random Forest, GBM= Gradient Boosting Machines, SGBM= Stochastic Gradient Boosting Machines

Detailed Example of Model Performance: First Grade Word Learning

The model comparisons contain information about how well each model explained the data, but not how interpretable or applicable the models are. First grade word learning data was used to demonstrate the differences in results produced by each model for this purpose. Each model was created for both decontextualized word learning and expressive labeling.

Multivariate Linear Regression

Multivariate linear regression results are presented in Table 5.7. Word frequency ($\beta = 8.04 \times 10^{-4}, p = 1.97 \times 10^{-3}$), age of acquisition ($\beta = -7.30 \times 10^{-2}, p < .001$), and level of concreteness ($\beta = 8.03 \times 10^{-2}, p < .01$) significantly predicted children's decontextualized word learning. These three lexical characteristics explained 69% of the variance in word learning (R^2= .69, F(5,137)= 60.97, p < .001). For expressive labeling, two

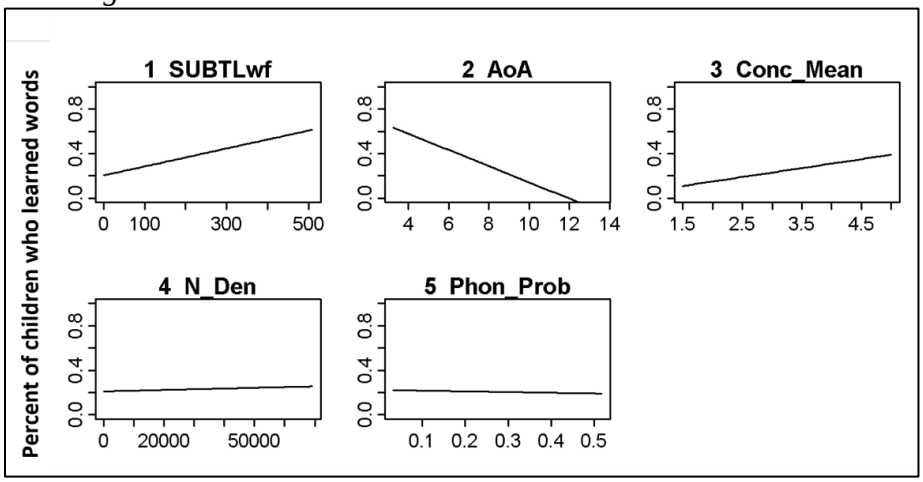
lexical characteristics were significantly predictive of expressive labeling, age of acquisition ($\beta = -5.72 \times 10^{-2}, p < .01$) and level of concreteness ($\beta = 1.55 \times 10^{-2}, p < .01$). These lexical characteristics explained 68% of the variance in expressive labeling ($R^2 = .68, F(5,137) = 58.28, p < .001$).

Table 5.7. *Multivariate Linear Regression Results for First Grade Word Learning*

	Decontextualized			Expressive		
	β	SE β	*p*-value	β	SE β	*p*-value
Intercept	.66	.12	< .001*	.47	.10	< .001*
Word Frequency	.0008	.0003	< .01*	.0004	.0002	.10
Age of Acquisition	- .073	.009	< .001*	- .0572	.0074	< .001*
Concreteness	.08	.02	< .001*	.09	.0155	< .001*
Neighborhood Density	6.43E-7	1.68E-6	.70	7.50E-7	1.50E-6	.61
Phonotactic Probability	- .0636	.12	.60	- .081	.11	.44
		$R^2 = .69$			$R^2 = .68$	

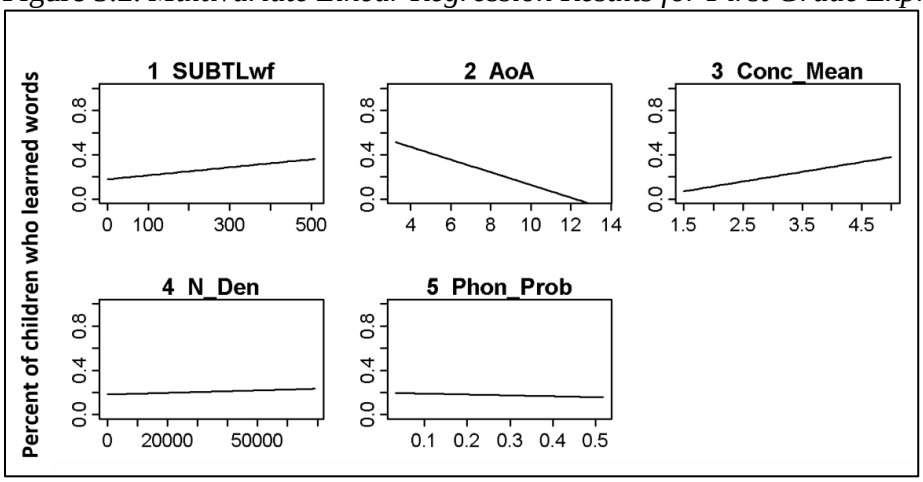
Trends for the individual predictor variables are displayed in Figures 5.1 and 5.2 with all other predictor variables held constant for models of decontextualized word learning and expressive labeling, respectively. For both regressions, word frequency and level of concreteness, are positively associated with children's word learning. Words with higher frequencies and words that are more concrete are easier for children to learn. The trend line for age of acquisition is negatively associated to word learning; words learned at a later age are more difficult for children to learn. Neighborhood density and phonotactic probability had a neutral effect on word learning.

Figure 5.1. *Multivariate Linear Regression Results for First Grade Decontextualized Word Learning*



Note. Scale for each x-axis differ based on lexical characteristic values.

Figure 5.2. *Multivariate Linear Regression Results for First Grade Expressive Word Learning*



Note. Scale for each x-axis differ based on lexical characteristic values.

Stepwise Regression

Neighborhood density and phonotactic probability were removed from the final models. Results are presented in Table 5.8. Age of acquisition ($\beta = -.07, p < .001$; $\beta = -5.87 \times 10^{-2}, p < .001$) and level of concreteness ($\beta = .08, p < .001$; $\beta = 8.83 \times 10^{-2}, p < .001$) significantly predicted children's decontextualized word learning and expressive labeling.

These predictors explained 68 – 69% of the variance in word learning ($R^2 = .69, F(3,139) = 102.5, p < .001$; $R^2 = .68, F(3,139) = 97.34, p < .001$).

Table 5.8. *Stepwise Regression Results for First Grade Word Learning*

	Decontextualized			Expressive		
	β	SE β	*p*-value	β	SE β	*p*-value
Intercept	.65	.11	<.001*	.47	.10	<.001*
Word Frequency	.0008	.0003	<.01*	.0004	.0002	.10
Age of Acquisition	- .07	.0083	<.001*	- .059	.0071	<.001*
Level of Concreteness	.08	.018	<.001*	.0883	.0154	<.001*
		R^2 = .69			R^2 = .68	

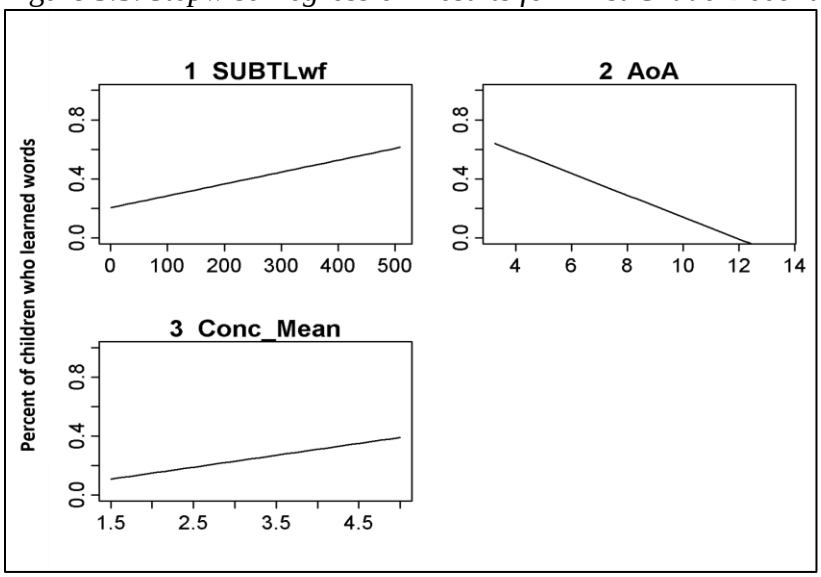
Figures 5.3 and 5.4 show the trends for each variable as all others are held constant for decontextualized and expressive learning, respectively. For both, word frequency and level of concreteness have a positive trend. Words with higher word frequencies and words that are more concrete were easier for children in first grade to learn. Age of acquisition had a negative trend, words with older ages of acquisition had lower rates of learning. Words learned later in childhood (e.g., AoA rating of 12 years old) were difficult for children in first grade to learn compared to words with earlier AoA ratings (e.g., 6 years old).

Ridge Regression

Figure 5.5 shows the ridge traces for the ridge regression models for decontextualized learning and expressive learning. As the ridge constant increases to the right, the variables shrink closer to zero. Using the ridge trace to find the optimal complexity parameter requires expertise to find a balance of stability and shrinkage. Variables that change the most have a higher variance, which ridge regression aims to limit by having higher shrinkage (Hoerl & Kennard, 1970; Friendly, 2013). Age of acquisition has the highest variance, followed by level of concreteness. Generally, the goal is to find where shrinkage has slowed, and variables are

visually stable. Visually, around 200 may be a good option but the vertical line represents the estimated location of k through analytic methods. Various analytic methods exist to determine the optimal shrinkage value (Gisela & Kibria, 2009) and minimization of the standard error of the cross-validation residuals was used for modeling the ILIAD data.

Figure 5.3. *Stepwise Regression Results for First Grade Decontextualized Word Learning*

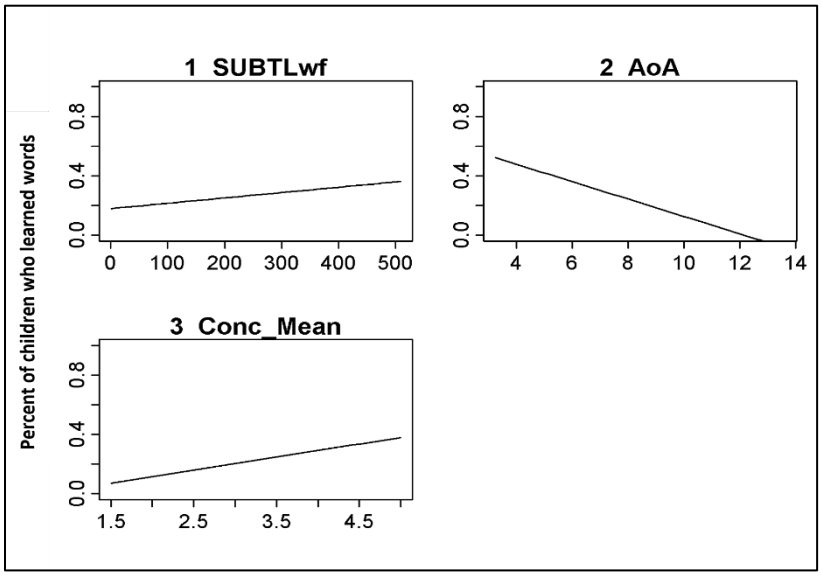


Note. Scale for each x-axis differs based on lexical characteristic.

Ridge regression results are presented in Table 5.9. Standard errors for the coefficients have not been including because they are misleading in penalized estimation methods (Casella et al, 2010). Standard errors are not very meaningful for strongly biased estimates and the penalized estimation procedure reduces variance in the estimators by introducing bias (Goesman et al, 2018). While significance is hard to directly measure, ridge regression is a penalized regression method that shrinks coefficients that do not contribute to explaining the response variable. Unlike LASSO, ridge regression cannot eliminate variables but only shrink them. For both decontextualized and expressive measures, neighborhood density ($\beta = 5.90 \times 10^{-7}, \beta = 7.20 \times 10^{-7}$) has been shrunk significantly from other lexical characteristics, followed by word

frequency ($\beta = 8.099 \times 10^{-4}, \beta = 3.90 \times 10^{-4}$). It can therefore be surmised that the the model found age of acquisition ($\beta = -6.69 \times 10^{-2}, \beta = -5.31 \times 10^{-2}$), level of concreteness ($\beta = 8.17 \times 10^{-2}, \beta = 8.63 \times 10^{-2}$), and phonotactic probability ($\beta = -8.71 \times 10^{-2}, \beta = -9.64 \times 10^{-2}$) to be the most significant lexical characteristics to influence word learning. These predictors explained 68-69% of the variance of word learning.

Figure 5.4. *Stepwise Regression Results for First Grade Expressive Word Labeling*



Note. Scale for each x-axis differs based on lexical characteristic values.

Figure 5.5. *Ridge Trace for Ridge Regression*

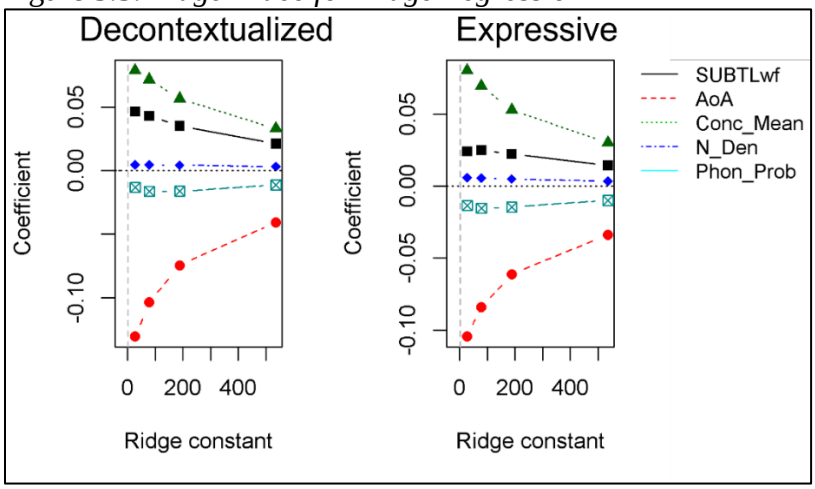
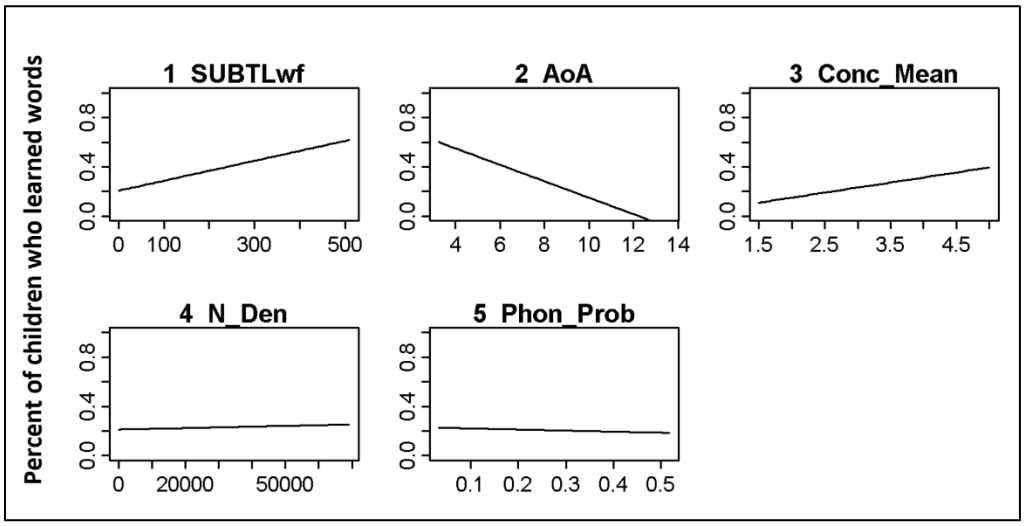


Table 5.9. *Ridge Regression Results for First Grade Word Learning*

	Decontextualized	Expressive
	β	β
Intercept	.605	.443
Word Frequency	8.099×10^{-4}	3.90×10^{-4}
Age of Acquisition	-6.69×10^{-2}	-5.31×10^{-2}
Concreteness	8.17×10^{-2}	8.63×10^{-2}
Neighborhood Density	5.90×10^{-7}	7.20×10^{-7}
Phonotactic Probability	-8.71×10^{-2}	-9.64×10^{-2}
	$R^2 = .69$	$R^2 = .68$

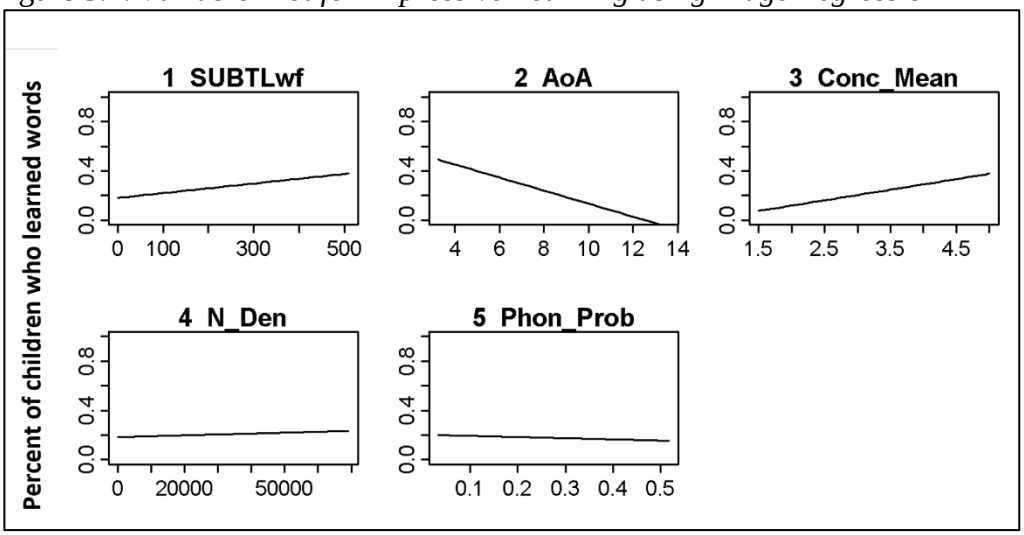
Trends for the individual predictor variables are displayed in Figures 5.6 and 5.7 with all other predictor variables held constant for models of decontextualized word learning and expressive labeling, respectively. For both regressions, word frequency and level of concreteness are positively associated with children's word learning. Words with higher frequency and that are more concrete are easier for children to learn, but word frequency has very little weight to word learning based on the shrinkage of the variable. Age of acquisition is negatively associated with children's word learn meaning that as the age of acquisition rating increases, a drop in learning is expected. Neighborhood density and phonotactic probability had a neutral effect on word learning, though neighborhood density has very little weight overall according to the models. These models follow closely with the results from multivariate linear regression.

Figure 5.6. *Variable Plot for Decontextualized Learning using Ridge Regression*



Note. Scale for each x-axis differs based on lexical characteristic values.

Figure 5.7. *Variable Plot for Expressive Learning using Ridge Regression*



Note. Scale for each x-axis differs based on lexical characteristic values.

Least Absolute Shrinkage and Selection Operator (LASSO)

The results for the regressions of decontextualized word learning and expressive labeling using LASSO are presented in Table 5.10. While LASSO shrinks or penalizes variables similar to ridge regression, LASSO can eliminate variables as well. Neighborhood density and

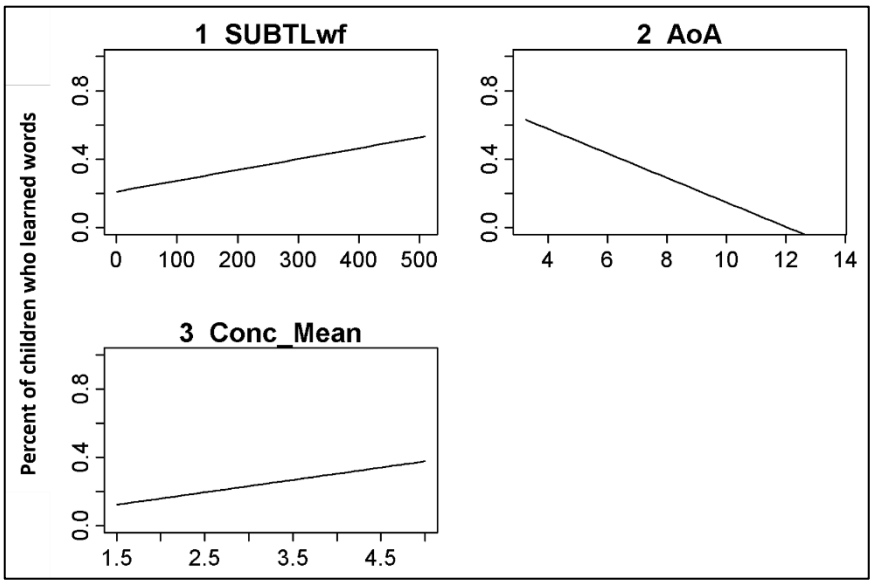
phonotactic probability have been eliminated from the final models. Significance values are inappropriate for LASSO but based on shrinkage and elimination, age of acquisition ($\beta = -6.97 \times 10^{-2}, \beta = -5.59 \times 10^{-2}$) and level of concreteness ($\beta = 6.65 \times 10^{-2}, \beta = 7.94 \times 10^{-2}$) are the most significant variables, followed by word frequency ($\beta = 5.07 \times 10^{-4}, \beta = 1.73 \times 10^{-4}$). According to these models, the lexical characteristics explained 68% of the variance in word learning.

Table 5.10. *LASSO Results for First Grade Word Learning*

	Decontextualized	Expressive
	β	β
Intercept	.663	.473
Word Frequency	5.07×10^{-4}	1.73×10^{-4}
Age of Acquisition	-6.97×10^{-2}	-5.59×10^{-2}
Concreteness	6.65×10^{-2}	7.94×10^{-2}
Neighborhood Density		
Phonotactic Probability		
	$R^2 = .68$	$R^2 = .68$

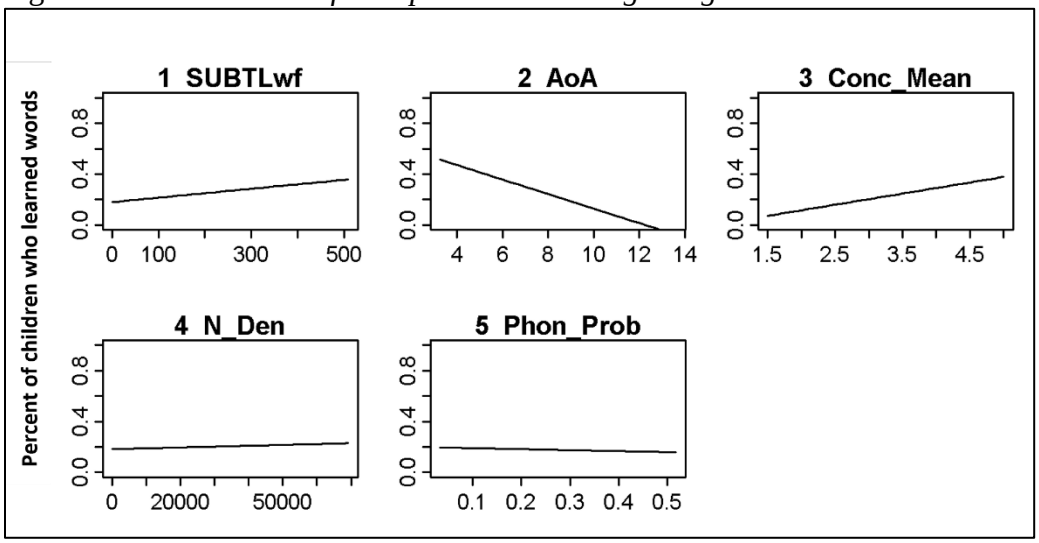
Figures 5.8 and 5.9 show the trends for each variable with all others held constant for decontextualized and expressive learning, respectively. Neighborhood density and phonotactic probability were eliminated from both models. Word frequency and level of concreteness both have a positive effect on word learning, though word frequency was shrunk giving it less weight. Age of acquisition has a negative trend, meaning that as the age of acquisition rating increases, it is expected that less children will learn the word at the first-grade level.

Figure 5.8. *Variable Plot for Decontextualized Learning using LASSO*



Note. Scale for each x-axis differs based on lexical characteristic values.

Figure 5.9. *Variable Plot for Expressive Learning using LASSO*



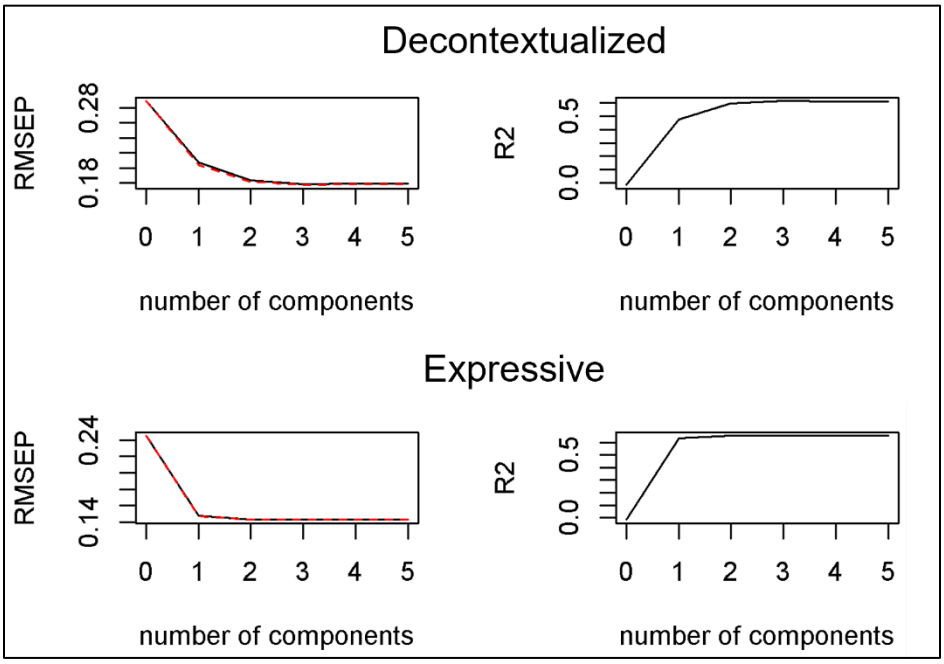
Note. Scale for each x-axis differs based on lexical characteristic values.

Partial Least Squares (PLS)

With partial least squares, a choice must be made for the number of components to project the variables onto. Figure 5.10 shows the impact the number of components has on the root mean square error and R^2 for decontextualized learning and expressive learning,

respectively. Based on the graphs, both models are optimal with two components, though 3 is also applicable for the expressive model. Analytic methods comparing the impact the number of components had on square error were used to confirm the selection for each model.

Figure 5.10. *Partial Least Squares Model Fit by Number of Components*



The results for the partial least squares regressions are presented in Table 5.11. As with other shrinkage methods, partial least squares shrinks the less impactful predictor variables. This is done by projecting the variables onto a latent structure that is not directly observed, by creating orthogonal score vectors by maximizing the covariance between different sets of variables (Rosipal & Krämer, 2005). For both models, age of acquisition ($\beta = -1.38 \times 10^{-1}, \beta = -1.13 \times 10^{-1}$) has the most influence on children's word learning. This is followed by level of concreteness ($\beta = 9.68 \times 10^{-2}, \beta = 9.57 \times 10^{-2}$) and word frequency ($\beta = 5.30 \times 10^{-2}, \beta = 2.58 \times 10^{-2}$). The smallest contributions are made by phonotactic probability ($\beta = -9.28 \times 10^{-3}, \beta = -1.05 \times 10^{-2}$) and neighborhood density

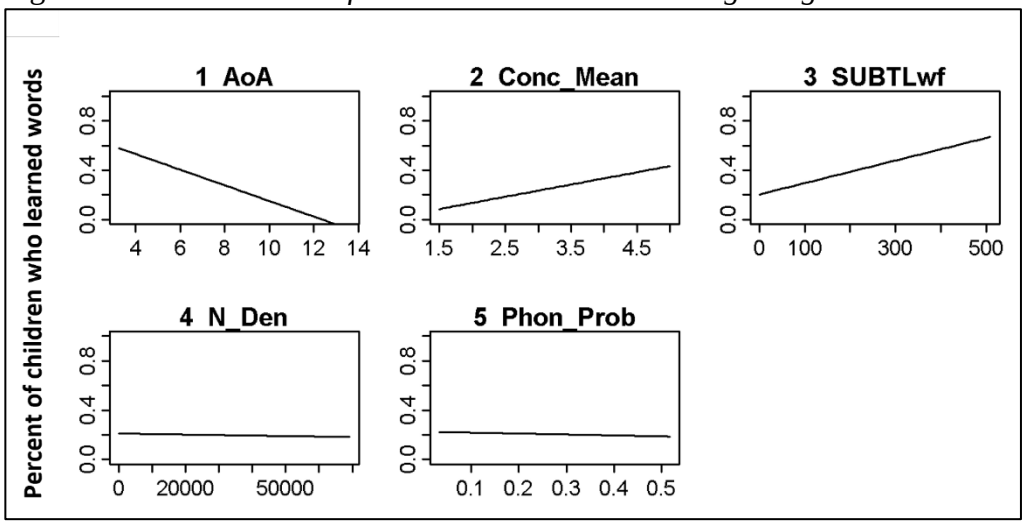
($\beta = -3.40 \times 10^{-3}, \beta = 7.88 \times 10^{-4}$). These lexical characteristics explained 68-69% of the variance in each model.

Table 5.11. *Partial Least Squares Regression Results for First Grade Word Learning*

	Decontextualized	Expressive
	β	β
Intercept	.520	.394
Word Frequency	5.30×10^{-2}	2.58×10^{-2}
Age of Acquisition	-1.38×10^{-1}	-1.13×10^{-1}
Concreteness	9.68×10^{-2}	9.57×10^{-2}
Neighborhood Density	-3.40×10^{-3}	7.88×10^{-4}
Phonotactic Probability	-9.28×10^{-3}	-1.05×10^{-2}
	$R^2 = .69$	$R^2 = .68$

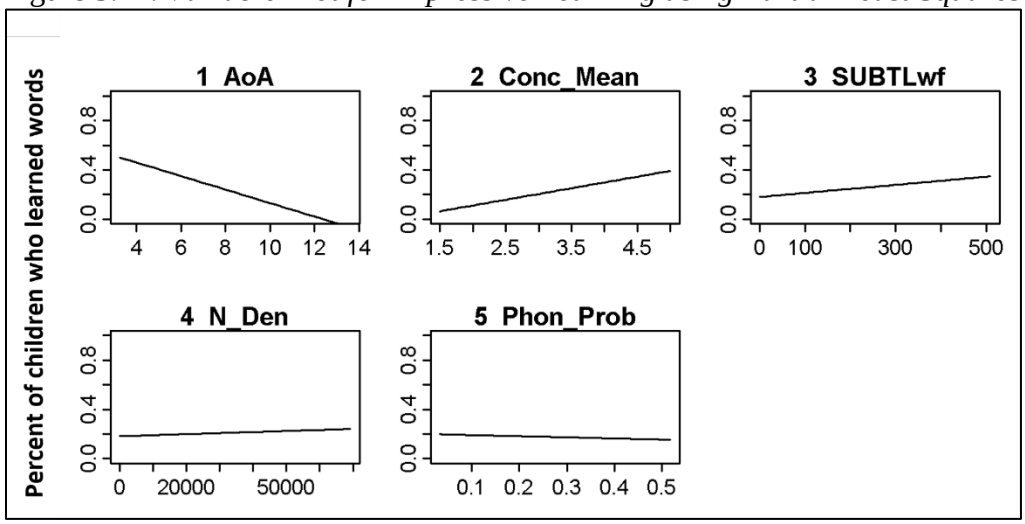
Figures 5.11 and 5.12 contain the trends for the individual predictor variables while all others are held constant for decontextualized learning and expressive tasks, respectively. Age of acquisition has a negative trend with first grader's word learning for both models. Predictors are ordered based on their importance, so age of acquisition has the largest impact, followed by level of concreteness, word frequency, neighborhood density, and phonotactic probability. Level of concreteness and word frequency have positive trends, meaning as the concreteness of a word increases and the word's frequency increases, higher word learning occurs. Neighborhood density has a positive trend and phonotactic probability a negative trend, but both have a smaller influence than other predictors.

Figure 5.11. *Variable Plot for Decontextualized Learning using Partial Least Squares*



Note. Scale for each x-axis differs based on lexical characteristic values.

Figure 5.12. *Variable Plot for Expressive Learning using Partial Least Squares*

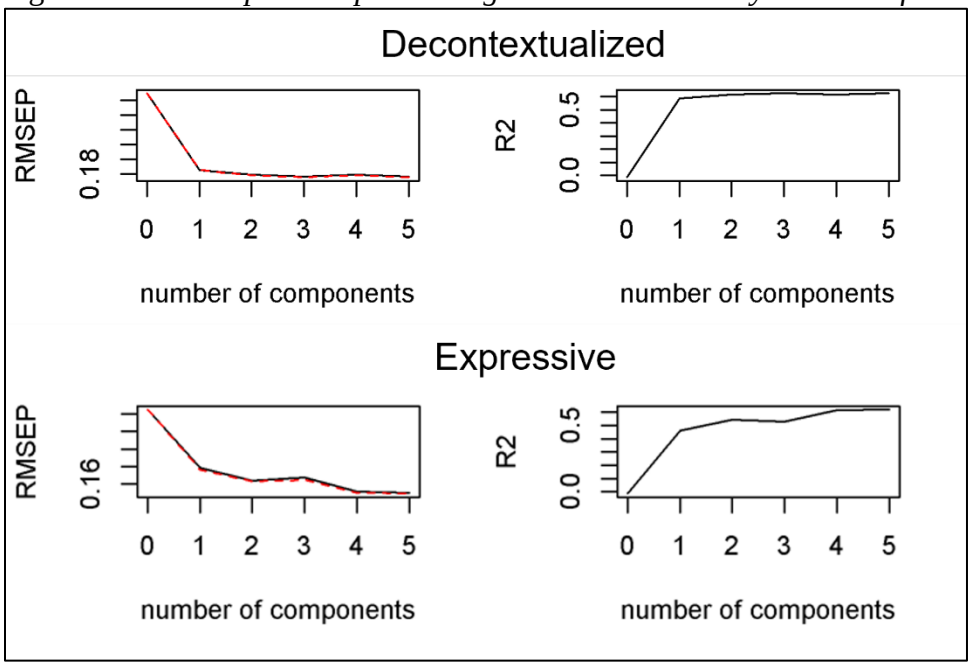


Note. Scale for each x-axis differs based on lexical characteristic values.

Principal Component Regression (PCR)

With principal component regression, a choice must be made for the number of latent factors to project the variables onto. Figure 5.13 shows the impact the number of components has on the root mean square error and R^2 for decontextualized learning and expressive learning. For the decontextualized model, one latent factor was used based on the graphs and analytic assessments. For expressive labeling, the graphs indicate three or four latent factors is optimal. Analytic methods indicate three latent factors is the best choice for this dataset.

Figure 5.13. *Principal Component Regression Model Fit by Number of Components*



Principal component regression results are presented in Table 5.12. PCR was chosen to analyze the ILIAD data because of the well-known collinear nature of lexical characteristics. It overcomes multicollinearity by using the principal components of the parameters as the regressors and selecting a subset of the latent factors. For decontextualized learning, one latent factor was used and for expressive labelling, three latent factors were used.

Table 5.12. *Principal Component Regression Results for First Grade Word Learning*

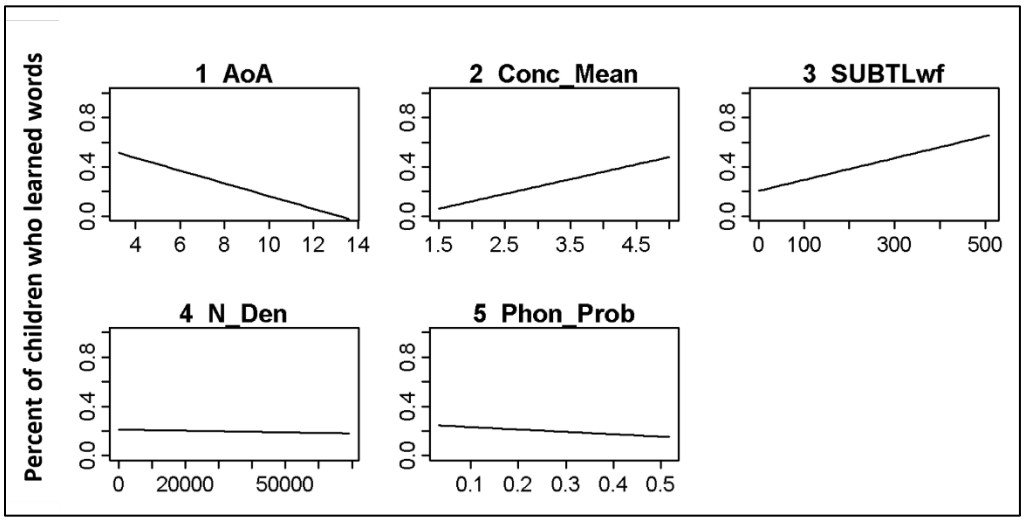
	Decontextualized	Expressive
	β	β
Intercept	.457	.311
Word Frequency	6.60×10^{-2}	2.48×10^{-2}
Age of Acquisition	-9.76×10^{-2}	-9.82×10^{-2}
Concreteness	8.98×10^{-2}	1.07×10^{-1}
Neighborhood Density	2.72×10^{-2}	3.25×10^{-3}
Phonotactic Probability	-5.89×10^{-2}	-1.66×10^{-2}
	$R^2 = .63$	$R^2 = .67$

Trends for the individual predictor variables are displayed in Figures 5.14 and 5.15 with all other predictor variables held constant for models of decontextualized word learning and expressive labeling, respectively. With the projection onto principal components, new insights can be seen from the graphs. Age of acquisition and phonotactic probability have negative trends while level of concreteness, word frequency, and neighborhood density have positive trend lines. What is special is the associated impact from lexical characteristics that have been smaller in other models. By considering latent space, we may be seeing the impact of lexical characteristics on word learning while limiting age of acquisition overwhelming the results. Because principal component regression is an unsupervised technique, we would need to do further analysis to determine more and validate the results.

Multivariate Adaptive Regression Splines (MARS)

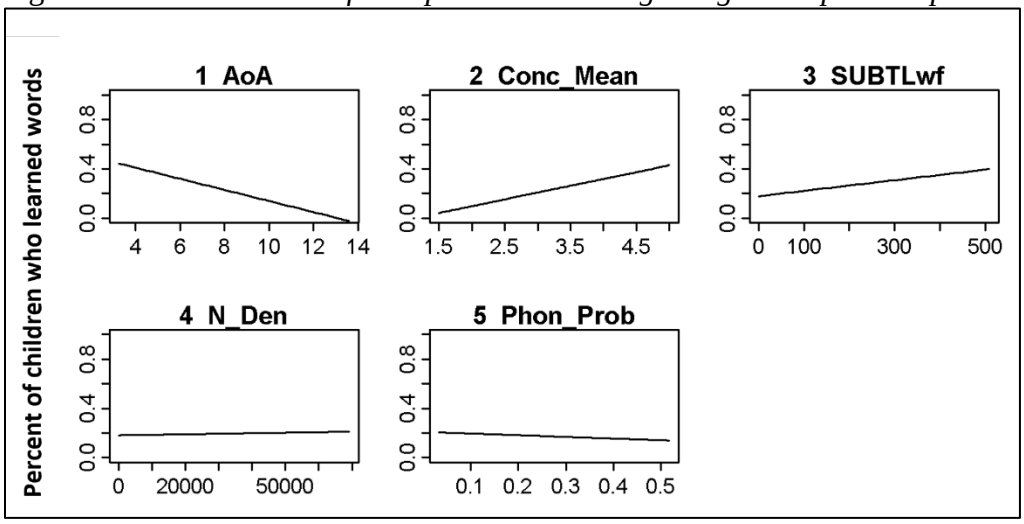
Because MARS uses recursive partitioning, variable selection is automatically considered during the backwards pass. Three criteria were used to determine variable importance: number of subsets (nsubsets), generalized cross validation (GCV), and residual sum of squares (RSS). Nsubsets is the number of model subsets that included that variable during the backward pass. During the backwards deletion step each variable was separately removed and the impact on the model was compared. This was repeated until ending at a model with one variable. MARS kept

Figure 5.14. *Variable Plot for Decontextualized Learning using Principal Component Regression*



Note. Scale for each x-axis differs based on lexical characteristic values.

Figure 5.15. *Variable Plot for Expressive Learning using Principal Component Regression*



Note. Scale for each x-axis differs based on lexical characteristic values.

track of each "best model" during each step and which variables were included. A variable included in more subsets was ranked higher in importance. Residual sum of squares was calculated for each subset relative to the prior subset during the backwards deletion phase. A score of 100 is always given to the most important variable and each preceding variable is scored

relative to that. The decreases in RSS for each variable across each subset that contained the variable are summed and ranked. The most important variable based on the summed decreases is scaled to 100. Generalized cross validation described in the model section works in a similar manner to RSS but using cross validation instead of RSS. The most important variable is scored as 100 and each preceding variable was ranked relative to that. Any of these criteria may be used as a threshold for inclusion in the final model.

By summarizing the number of subsets, GCV, and RSS values for variable selection, variables are ranked by the level of importance. As illustrated in Table 5.13, all variables were included in the final model for decontextualized word learning. The lexical characteristics in order of importance were age of acquisition, level of concreteness, neighborhood density, word frequency, and phonotactic probability. In contrast, variable selection for the final model for expressive learning removed neighborhood density, word frequency, and phonotactic probability. The remaining variables ranked by importance were age of acquisition and level of concreteness.

Table 5.13. *MARS Variable Selection for first Grade Word Learning*

	Decontextualized			Expressive		
Variable	nsubsets	GCV	RSS	nsubsets	GCV	RSS
Age of Acquisition	7	100	100	4	100	100
Level of Concreteness	5	22.1	26	3	27	28.5
Neighborhood Density	4	16	20.3	0	0	0
Word Frequency	2	8.1	12.1	0	0	0
Phonotactic Probability	1	5.5	8.3	0	0	0

Results for MARS is presented in Table 5.14. For decontextualized word learning and expressive labeling, the MARS analyses selected eight basis functions and five basis functions, respectively. The basis functions for each variable indicate hinge location, whether the regression coefficient is to the right or left of the hinge, and the direction of the local trend (positive or negative). Each hinge acts a local interval boundary within the global model. For example, in the

decontextualized model, age of acquisition has hinges at 5.37, 7.81, and 8.45. During the pruning process (backwards step), some local regressions were removed because other variables better-predicted word learning for that global region. The predictor variables explained 84 – 86% of the variance in decontextualized word learning ($R^2 = .84$) and expressive labeling ($R^2 = .86$).

Table 5.14. *MARS Results for First Grade Word Learning*

Decontextualized				Expressive			
Predictor	Type	Hinge Location	Coefficient	Predictor	Type	Hinge Location	Coefficient
(Intercept)			0.91	(Intercept)			0.14
Word Freq	Left	32.22	-0.003	AoA	Left	7.33	0.16
AoA	Right	5.37	-0.27	AoA	Right	7.33	-0.015
AoA	Right	7.81	0.42	Concrete	Right	2.55	0.05
AoA	Right	8.45	-0.17	Concrete	Right	3.97	0.18
Concrete	Right	3.00	0.07				
N Den	Left	126.04	-0.001				
Phono Prob	Left	0.08	-3.05				

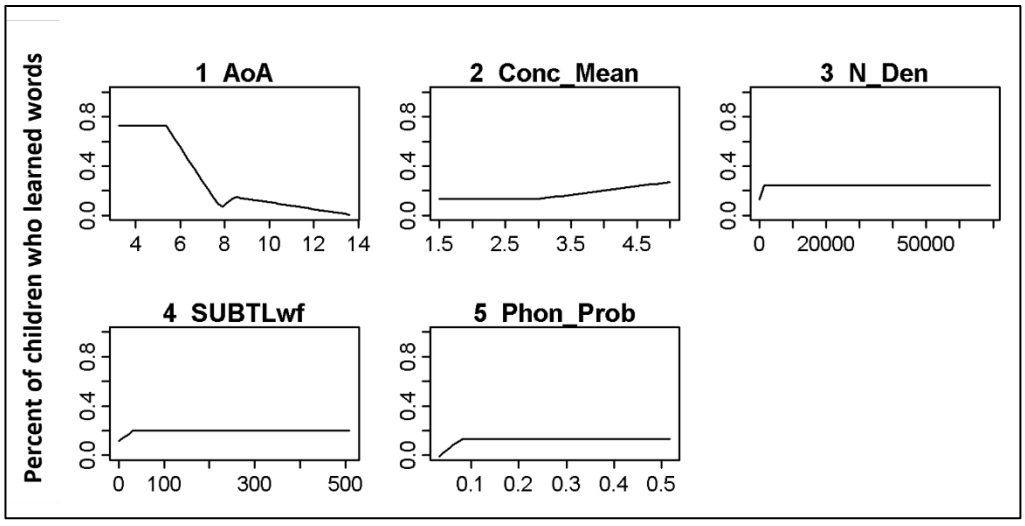
Note. Word Freq= Word Frequency; AoA= Age of Acquisition; Concrete= Level of Concreteness; N Den= Neighborhood Density; Phono Prob= Phonotactic Probability

It is difficult to visualize a higher dimensional model, so the above regression may be difficult to interpret. The models for decontextualized and expressive learning are composed of the predictor variable, the location of a hinge (spline), and the coefficient for the local regression within the variable and its location relative to the hinge (i.e., type). For example, in the decontextualized model for first grade learning, word frequency has a hinge at 32.22 with a coefficient of -.003 to the left of the hinge. To understand the relationship between significant lexical characteristics and children's word learning, individual variable plots were used. Figures 5.16 and 5.17 show the regression of each variable with all other variables held constant for both decontextualized word learning and expressive labeling. Graphical representations of multiple and stepwise regressions are represented by static linear relations that have an overall positive, negative, or neutral trend. With MARS, a more precise relationship is represented because hinges

created local intervals. These interval trends can vary from hinge to hinge allowing for a dynamic representation of relations among word learning and lexical characteristics.

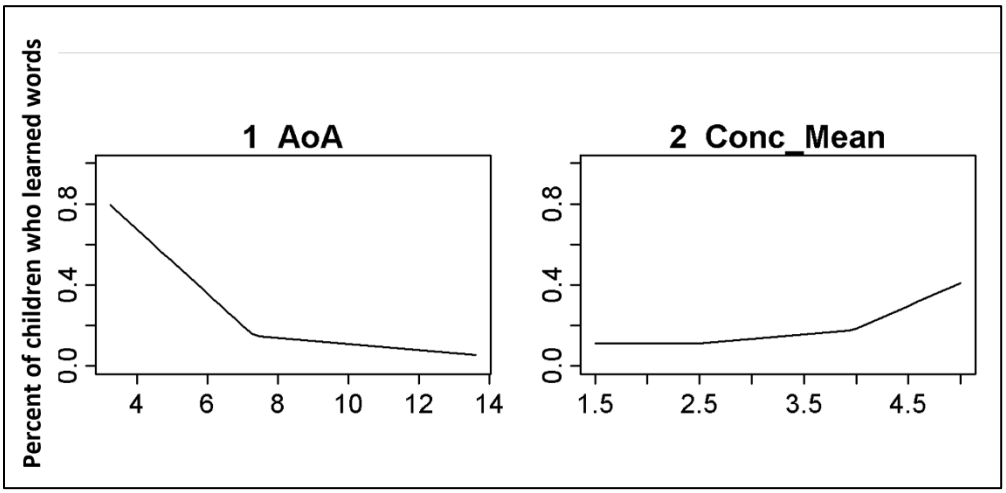
For example, in the model of decontextualized word learning, age of acquisition had negative and positive associations with word learning. There was a slow downward trend until the hinge at 5.37 meaning that 80% of children learned words with an age of acquisition starting at 2.6. From this point learning decreased as AoA increased. Learning dropped from approximately 70% to 15% in the next interval between hinges at 5.37 and 7.81 years old. A small upward trend showed that learning increased from 15% to 20% for words with AoA ratings ranging from 7.81 to 8.45. Finally, another slow decrease in learning occurred for words with AoA ratings older than 8 and a half years.

Figure 5.16. *Variable Plot for Decontextualize Learning using MARS*



Note. Scale for each x-axis differs based on lexical characteristic values.

Figure 5.17. *Variable Plot for Expressive Learning using MARS*



Note. Scale for each x-axis differs based on lexical characteristic values.

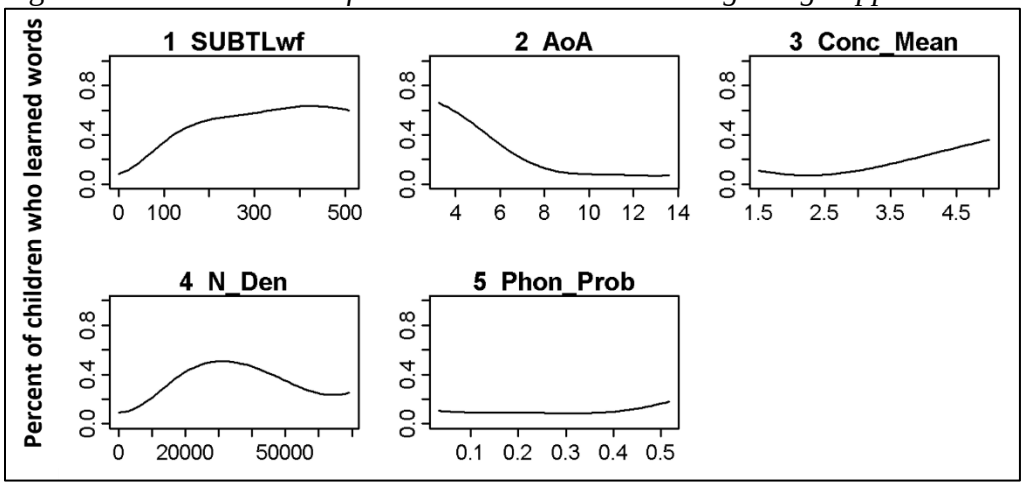
When compared to the decontextualized model, age of acquisition in the model of expressive learning had an overall negative association with word learning with fewer hinges. Age of acquisition had a rapid downward trend until the hinge at 7.33 years old. The percentage of children learning words dropped from 80% to approximately 15% as age acquisition ratings increased from 3.25 to 7.33 years old. After this interval, the rate of learning still decreases, but at a slower rate, from 15% to 0% of children learning words with an AoA rating older than 7.33 years old. A step-by-step guide modeling with MARS can be found in Appendix I.

Support Vector Regression (SVR)

The parameter influence with all other parameters held constant for support vector regression for decontextualized learning and expressive labeling can be found in Figures 5.18 and 5.19, respectively. Prior distribution data was not known, so a grid search was performed across parameters and kernels to determine the optimal model for the data. A radial basis function was used as the kernel for decontextualized and expressive models. Hyperparameters were also found at this step for decontextualized learning ($\varepsilon = 0.1, \gamma = 0.2, c = 2, sv = 117$) and expressive learning ($\varepsilon = 0.1, \gamma = 0.2, c = 3, sv = 115$), where ε is the margin of

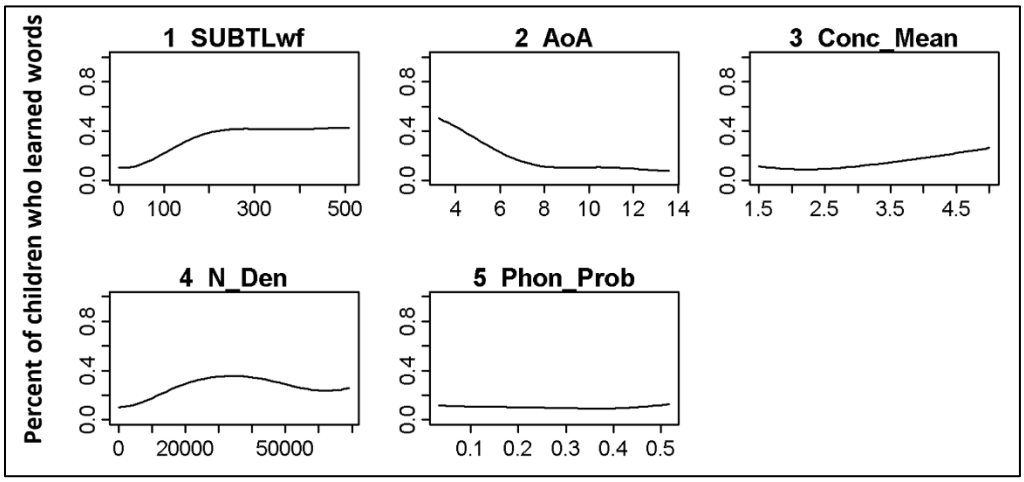
tolerance, γ controls soft margins, c is the cost parameter, sv is the number of support vectors used in the model. There is no traditional regression output for support vector regressions, so it is important to consider what the model shows for the individual influence of each variable. For both models, word learning increases positively with word frequency until it hits a threshold where gains become neutral. This is around 400 for decontextualized learning and 200 for expressive labeling. As the age of acquisition rating of words increases, the word learning decreases. This occurs steadily until a rating of 8 and then becomes neutral, likely because nearly no words are being learned by first graders with higher ratings. Level of concreteness has a positive trend with word learning after a concreteness score of 2.5. Neighborhood density has a positive trend with word learning until a rate of 30,000. After that, the word learning drops as neighborhood density scores increase. Phonotactic probability does not appear to have much impact on word learning and is neutral with a small positive trend after a phonotactic probability of 0.4.

Figure 5.18. *Variable Plot for Decontextualized Learning using Support Vector Regression*



Note. Scale for each x-axis differs based on lexical characteristic values.

Figure 5.19. *Variable Plot for Expressive Learning using Support Vector Regression*

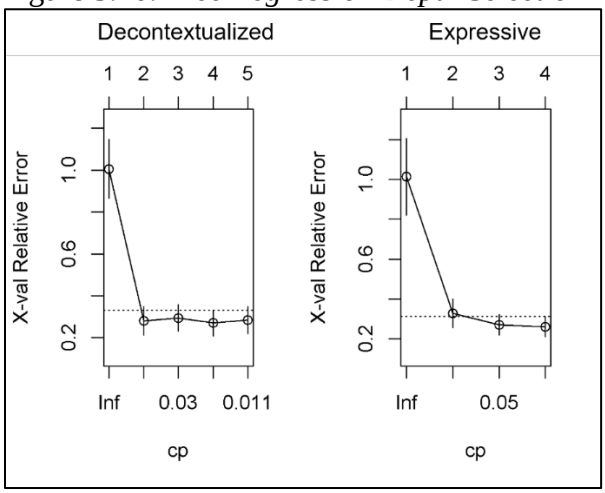


Note. Scale for each x-axis differs based on lexical characteristic values.

Regression Trees

In order to limit overfitting, pruning is often required for tree regression. The optimal tree depth can be determined by calculating the complexity parameter that minimizes error while keeping the tree depth minimal. Figure 5.20 display the impact of the complexity parameter on the error rate and its associate tree depth for decontextualized learning and expressive labeling, respectively. The complexity parameters were found analytically, giving final model for the tree the regress the predictors.

Figure 5.20. *Tree Regression Depth Selection*



Figures 5.21 and 5.22 shows the dendrogram for the pruned tree modeling decontextualized word learning and the associated variable plots, with all other variables held constant. The two variables it chose to keep are age of acquisition and word frequency. During each node, the split that minimized RSS the most was chosen. The first split is at age of acquisition rating of 6.725. If the word has a higher age of acquisition, it follows the left branch to the next node. If it is less than 6.725 it follows the right path. On the left path, it again splits based on age of acquisition, this time at 9.225. If the word has a higher rating, it follows the left path to a terminal node. If it is less, it follows the right path to a node that splits based on word frequency of 25.2. These branches can be seen in the variable plot, where the word frequency split at 25.2 and higher age of acquisition lowering children's word learning.

Figure 5.21. *Tree Diagram for Decontextualized Learning*

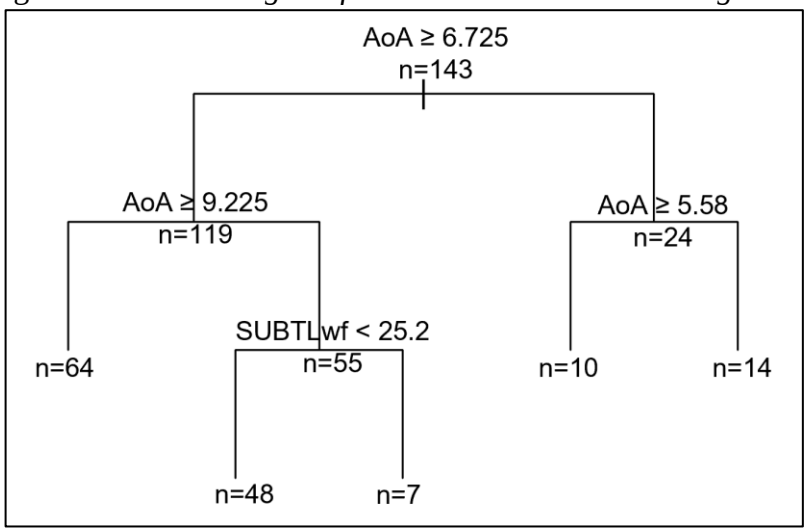
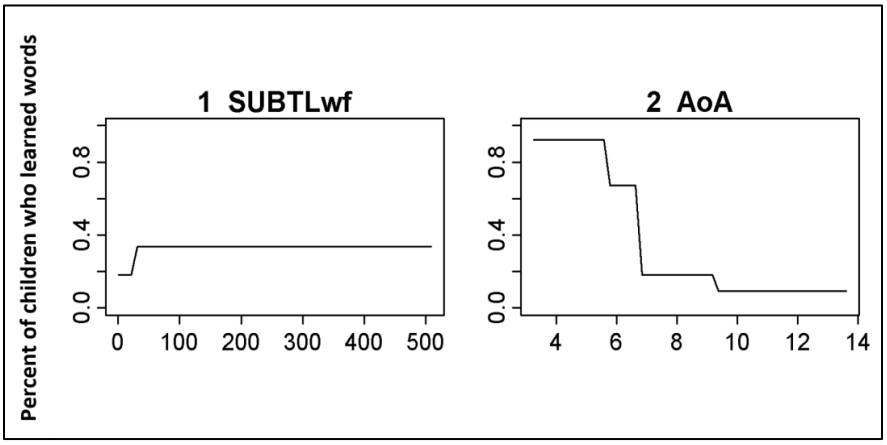


Figure 5.22. *Variable Plot for Decontextualized Learning using Regression Trees*



Note. Scale for each x-axis differs based on lexical characteristic values.

Figures 5.23 and 5.24 display the expressive tree model dendrogram and variable plots respectively. This model determined two variables explain word learning, age of acquisition and level of concreteness. The first split occurs at age of acquisition rating 6.57, with greater than or equal values taking the left path. At the resulting node, a split occurs at level of concreteness 2.845. These splits can be seen in the associated predictor plots with one positive shift for level of concreteness at 2.845 and drops in word learning at each split as age of acquisition increases.

Figure 5.23. *Tree Diagram for Exppressive Labeling*

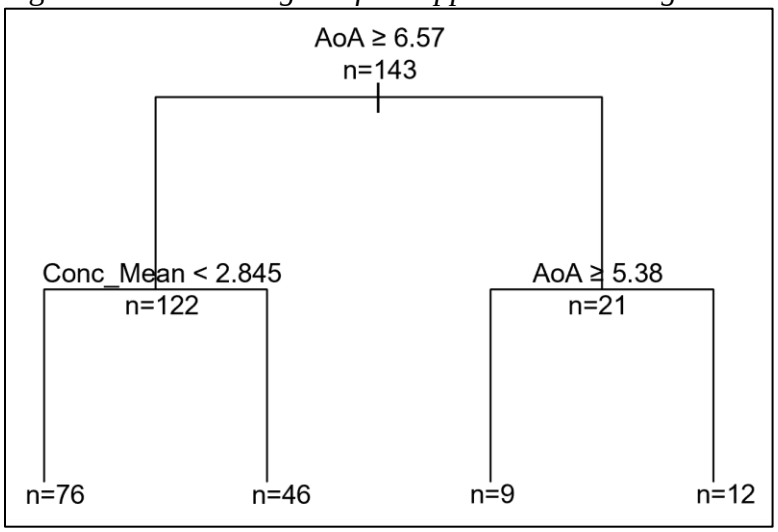
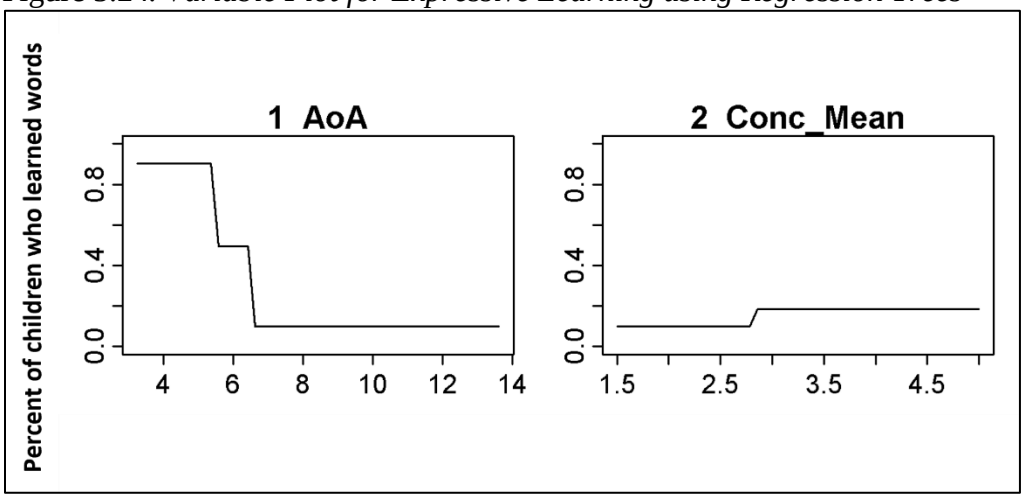


Figure 5.24. *Variable Plot for Expressive Learning using Regression Trees*



Note. Scale for each x-axis differs based on lexical characteristic values.

Random Forest

Figure 5.25 shows the error rate based on the number of trees created for random forest models for decontextualized learning and expressive labeling respectively. From the decontextualized figure, the error rate begins around 0.5 for error with one tree and the error rate increases as more trees are added. Once a threshold of around 15 trees is reached, error begins to quickly lower as more trees are added. The error rate continues to decrease as more trees are included at a slower rate. Each model was run with 500 trees, though comparable error was possibly with 200-300 trees.

Because of the ensemble nature of random forest regression, there is no simple model to consider. Instead, there are many regression trees with an aggregate vote for the outcome based on the lexical characteristics. This allows for precise results because every tree has equal impact on the vote and the random nature of choosing predictors when creating the trees can help eliminate noise and the impact of outliers. Figures 5.26 and 5.27 show the trends for each lexical characteristic's impact on word learning with other predictors held constant. For both models

word frequency increases positively with word learning until around 80 for decontextualized learning and 280 for expressive labeling, then plateaus. Age of acquisition has a negative trend in relation to word learning, with the largest drop between age ratings of 6 and 8. First graders are generally 6 years old, so this shows that words above their age rating become more difficult to learn at their age. Level of concreteness has a positive trend with word learning for both models, increasing more quickly as concreteness increases. Neighborhood density agrees with other models that it increases until around 25 and then plateaus for decontextualized learning and slowly increases for expressive labeling. Phonotactic probability appears to be neutral, and the movement is likely due to noise.

Figure 5.25. *Optimal Number of Trees for Random Forest*

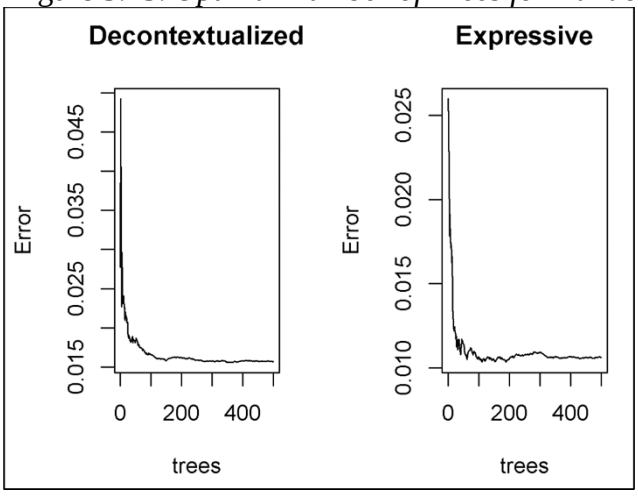
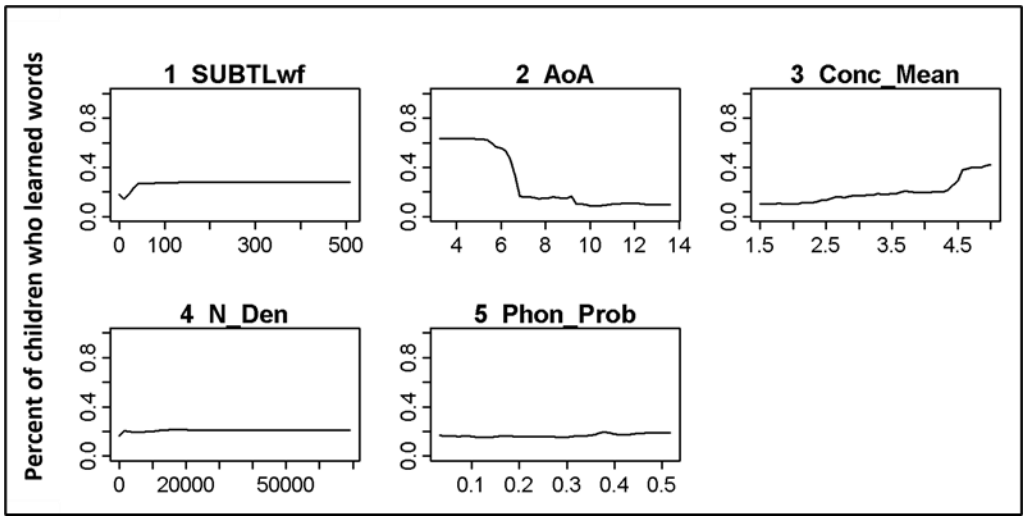
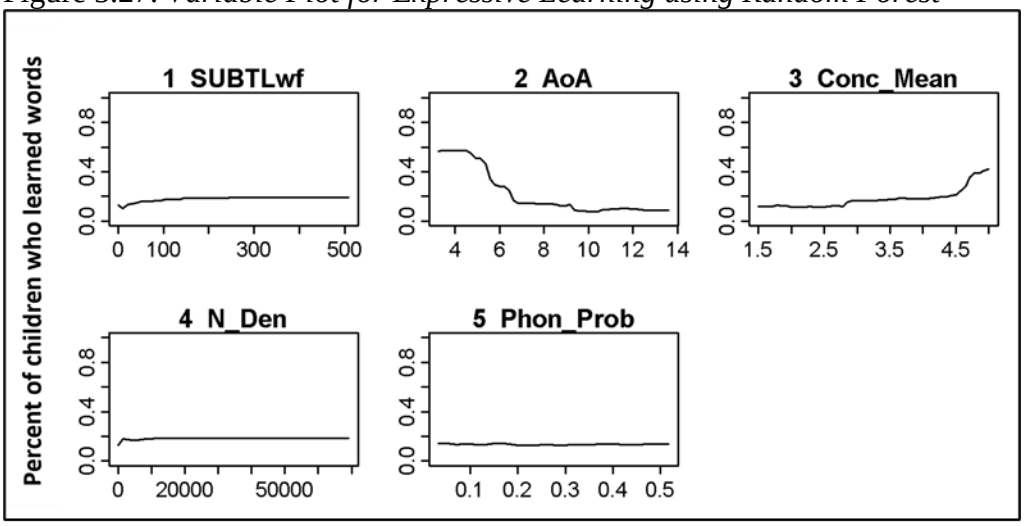


Figure 5.26. *Variable Plot for Decontextualized Learning using Random Forest*



Note. Scale for each x-axis differs based on lexical characteristic values.

Figure 5.27. *Variable Plot for Expressive Learning using Random Forest*



Note. Scale for each x-axis differs based on lexical characteristic values.

Gradient Boosting Machines (GBM)

Figure 5.28 displays the variable importance according to the gradient boosting machines for decontextualized learning and expressive labeling. For both models, age of acquisition is the variable with the largest impact on explaining word learning. Following age of acquisition level of concreteness is a distant second most important lexical characteristic for decontextualized

learning and a closer second for expressive labeling. Neighborhood density, word frequency, and phonotactic probability follow but do little explain word learning for the ILIAD data.

Trends for the individual predictor variables are displayed in Figures 5.29 and 5.30 with all other predictor variables held constant for models of decontextualized word learning and expressive labeling, respectively. Gradient boosting machines can perform variable selection based on the trees built during its creation and the graphs are ordered by importance allowing the focus to be on parameters that impact word learning most. Age of acquisition has a pronounced drop starting just before the age of 6 years until around 7 years. This supports what would be expected based on the ages of the students in first grade. Level of concreteness has a positive trend that rapidly increases between a rating of 4 and 4.5. Word frequency, neighborhood density, and phonotactic probability have a small upward trend but are mostly neutral.

Stochastic Gradient Boosting Machines

Figure 5.31 displays the variable importance according to the stochastic gradient boosting machines for decontextualized learning and expressive labeling. For both models, age of acquisition was the variable with the largest impact on explaining word learning followed distantly by level of concreteness. Neighborhood density, word frequency, and phonotactic probability follow but do little explain word learning for the ILIAD data.

Trends for the individual predictor variables are displayed in Figures 5.32 and 5.33 with all other predictor variables held constant for models of decontextualized word learning and expressive labeling, respectively. The iterative learning of stochastic gradient boosting machines allows the model to focus on the predictors with the largest impact. Because of this, age of acquisition has a pronounced drop starting just before the age of 6 years until around 7 years. This is a strong indication that the age of acquisition lexical characteristic is incredibly precise at predicting word

learning. Level of concreteness has a slight positive trend until it has a quick increase between a rating of 4 and 4.5. Word frequency, neighborhood density, and phonotactic probability have a small upward trend but are mostly neutral.

Figure 5.28. *Variable Importance for Gradient Boosting Machines*

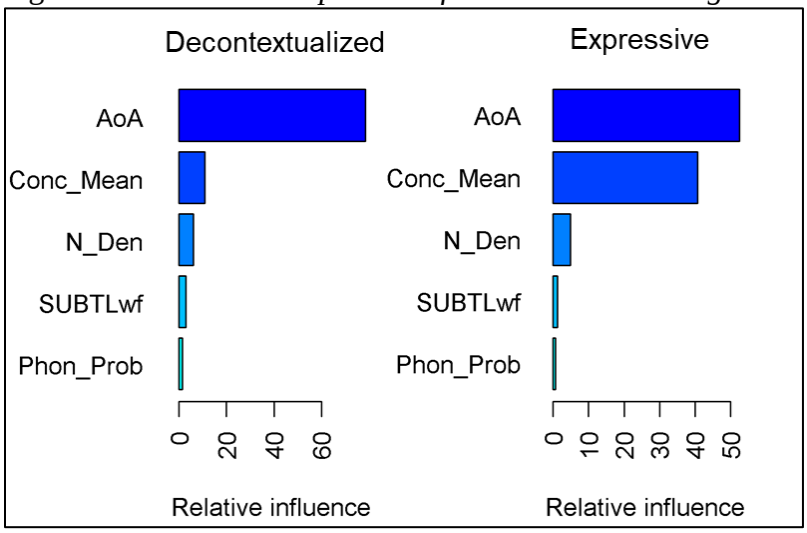
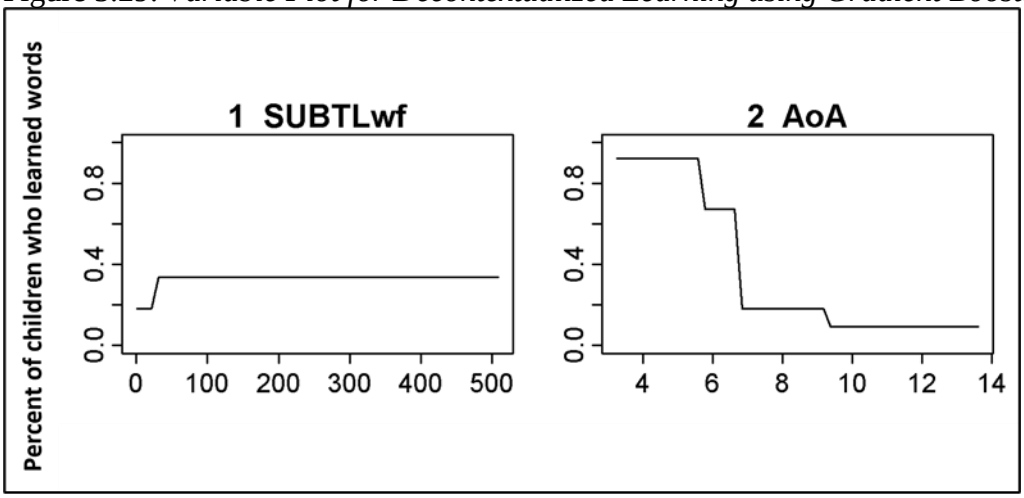
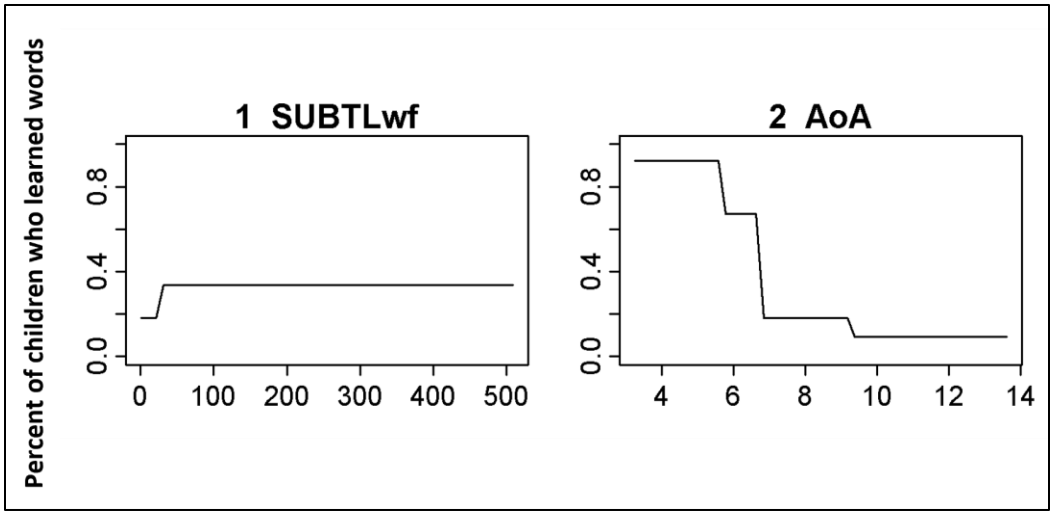


Figure 5.29. *Variable Plot for Decontextualized Learning using Gradient Boosting Machines*



Note. Scale for each x-axis differs based on lexical characteristic values.

Figure 5.30. *Variable Plot for Expressive Learning using Gradient Boosting Machines*



Note. Scale for each x-axis differs based on lexical characteristic values.

Figure 5.31. *Variable Importance for Stochastic Gradient Boosting Machines*

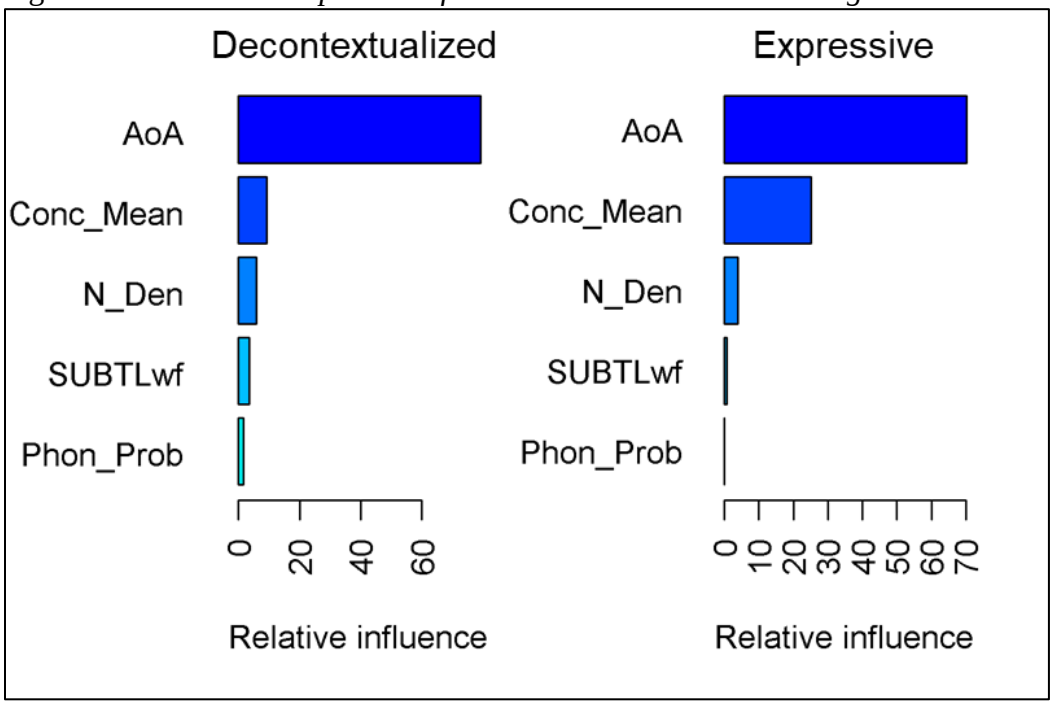
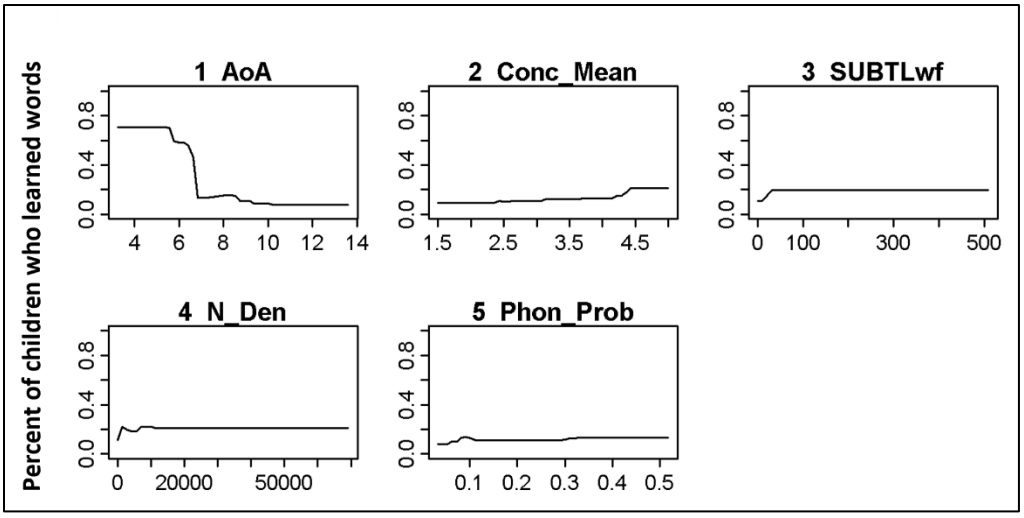
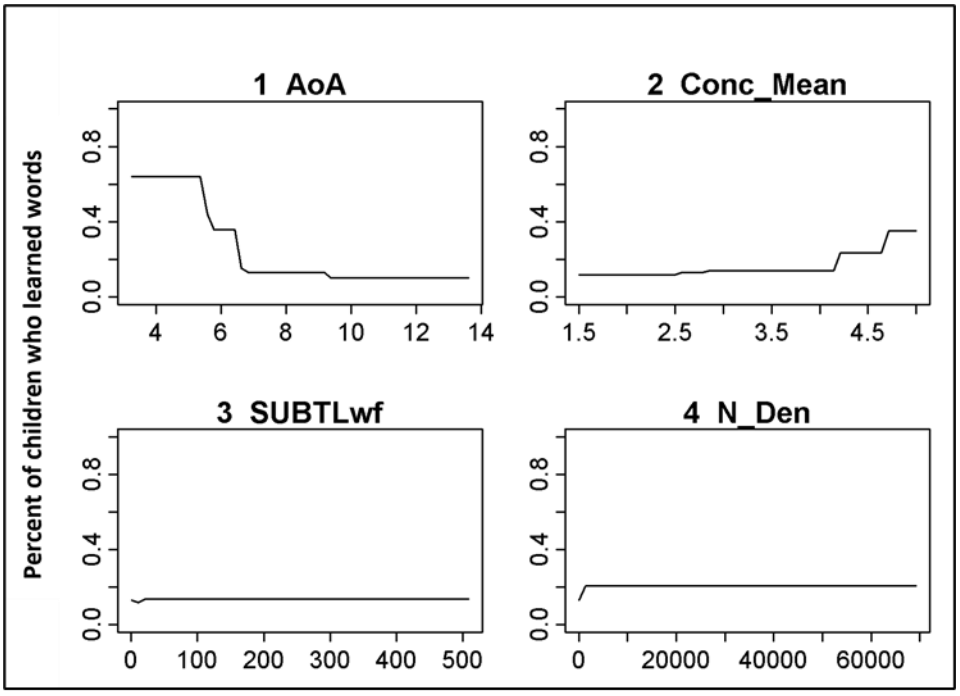


Figure 5.32. *Variable Plot for Decontextualized Learning using Stochastic Gradient Boosting Machines*



Note. Scale for each x-axis differs based on lexical characteristic values.

Figure 5.33. *Variable Plot for Expressive Learning using Stochastic Gradient Boosting Machines*



Note. Scale for each x-axis differs based on lexical characteristic values.

Discussion

Data from first grade was used to demonstrate the differences in results produced by the three modeling methods and supports the argument for adopting MARS as an alternative modeling method as opposed to other simpler regression models. Each model presented useful information used to make conclusions about the impact of the relevant lexical characteristics on word learning for the ILIAD study. Regression analyses revealed different relevant predictors of word learning. Multivariate linear regression analysis indicated word frequency, age of acquisition, and level of concreteness were significantly related to both decontextualized word learning and expressive labeling. This model accounted for 68 – 69% of the variance in word learning. It is important to note that insignificant variables remained in the model. Often the simplest model that accurately describes the data is most desired. In the case of our data, multivariate linear regression did not create the simplest model because it included extraneous variables. Variable selection methods are used to compensate for this.

Stepwise regression is a variable selection method that attempts to find the optimal model. Results for this analysis indicated word frequency, age of acquisition, and level of concreteness were significant predictors of decontextualized word learning and dropped both neighborhood density and phonotactic probability from the model completely. For expressive learning, the model the same variables, but word frequency was not a significant predictor of word learning. This model accounted for 68 – 69% of variance in word learning and demonstrated positive associations between word learning and word frequency and level of concreteness, and a negative association between age of acquisition and learning. Words that occur more frequently, had higher ratings of concreteness (words that are more concrete) and words with lower age of acquisition (words learned earlier) were easier for first graders to learn.

This is all we can say about lexical characteristics and word learning because that is all this analysis is telling us, anything beyond this would be speculative at best. A more sophisticated modeling method is needed to further investigate the nuanced relationship between lexical characteristics and word learning.

Ridge regression and LASSO performed equally well by the metrics MAE, MSE, RMSE, and R^2 in most instances. For the full models and first-grade models, both decontextualized learning and expressive learning, they performed equally well. For the second-grade model, they performed the same for expressive labeling but LASSO outperformed ridge regression by a small amount for each metric. For the third-grade model, LASSO had much higher R^2, while still small overall, but the error measures were similar for both decontextualized learning and expressive labeling.

For the first-grade example, both methods performed the same according to the error measures and R^2. Looking at the ridge regression, it shrunk the neighborhood density parameter significantly, as well as word frequency for both decontextualized and expressive learning outcomes. Ridge cannot eliminate variables, so it may shrink some until they are insignificant to the model. Based on this, ridge had the most weight associated with the parameters age of acquisition, level of concreteness, and phonotactic probability. LASSO, on the other hand, can eliminate variables from the regression. For both decontextualized and expressive outcomes, it eliminated neighborhood density and phonotactic probability, and assigned word frequency a smaller weight by shrinking the coefficient relative to the others. The regression found age of acquisition and level of concreteness to be the most impactful of the lexical characteristics on word learning, followed by word frequency.

Both ridge regression and LASSO generally performed equally for the ILIAD data, especially first grade. Their performance was equivalent to multivariate linear regression and stepwise regression outcomes for the full models, first grade models, and second grade models. Both had similar error measures to multivariate linear regression and stepwise regression for third grade but explained less of the variance (R^2). This means that they shrinkage did not have much of an impact on understanding the data. Both ridge regression and LASSO are linear methods that were chosen to account for the multicollinearity, and this demonstrates that while some multicollinearity exists, it did not have much impact on model building.

Similarly, partial least squares and principal component regression performed equally to ridge regression and each other on most datasets. For first grade decontextualized and third grade decontextualized and expressive models, they performed similarly for error metrics but PLS explained more of the variance than PCR. The performance of these two shrinkage regression methods supports that the multicollinearity does not have a very strong influence on modeling the data. PLS and PCR would have been stronger candidates with more variables and were not appropriate for the ILIAD dataset. Overall, shrinkage methods performed equally or worse than multivariate linear regression.

MARS is designed to adaptively use a combination of basis functions and hinges to balance precision with simplicity to create robust models. MARS error metrics were slightly better than the error metrics of multivariate linear regression and stepwise regression. However, MARS explained more of the variance in word learning $\left(R^2_{Decon} = .84, R^2_{Express} = .86\right)$ than multivariate linear regression and stepwise regression $\left(R^2_{Decon} = .69, R^2_{Express} = .68\right)$. MARS provided nuanced information about the predictive nature of lexical characteristics on word learning using information based on hinges and local intervals. These intervals are adaptive to

variability in learning and can change as a function of the predictors. Hinge placement creates local regressions that combine to form a complete regression model and these localized trends may vary for each lexical characteristic.

Support vector regression is a flexible method for modeling a variety of data based on the selection of a kernel. It outperformed all the shrinkage methods considered based on every error measure and R^2 for all models. For the second-grade model based on expressive tasks, it explained more of the variance than any other model. SVR is an adaptable method that can work with nonlinear data. This gives a better picture of what the impact lexical characteristics have on children's word learning. For age of acquisition, instead of steadily dropping we can see that word learning drops sharply between AoA ratings of 4 and 8 and because neutral. This may support that words become more difficult for children to learn as the rating goes up until a threshold where it becomes too difficult in general for children in first grade. Word frequency influences word learning sharply in a positive way and then becomes neutral at a frequency score of 400 for decontextualized word learning and 200 for expressive labeling. Higher word frequency leads to higher word learning up until a point, but then ceases to improve. These types of insights can give researchers a better, more nuanced view of how lexical characteristics influence word learning.

Support vector regressions did as well as MARS by each metric for all models. This supports that it is a strong alternative candidate for describing the influence of lexical characteristics on word learning. The weakness of SVR as a model is that it does not have an easy to interpret regression like prior methods that have been considered. Because of the nature of practitioners in the field, it may be hard to interpret the results because of the "black box" nature of the model. Support vector regression relies on the parameter's proximity to a

hyperplane and support vectors and may be difficult for researchers to comprehend. MARS performed equally well and is more interpretable, including having a more traditional regression model.

Regression trees performed nearly as well as SVR for all models, other than the first grade expressive model where it did slightly better. Based on the splits chosen, the model performs variable selection and for the first-grade example only age of acquisition and level of concreteness were included for both decontextualized and expressive tasks. While the method did well overall by the metrics of comparison, its discrete nature does not make it as applicable in practice. The dendrograms for the model can assist researchers in understanding the outcomes for each set of lexical characteristics, which could be useful for building a word learning decision framework for future studies.

Random forests, gradient boosting machines, and stochastic gradient boosting machines were the tree-based ensemble methods considered. Overall, they outperformed the other models including MARS. Random forest had the best error measures for $\frac{7}{8}$ of the models with GBM and stochastic GBM outperforming it for the third-grade decontextualized model. Gradient boosting and stochastic gradient boosting explained the highest amount of the variance in $\frac{7}{8}$ of the models, with SVR outperforming both for the second-grade expressive model. As with many of the models, random forest had a low R^2 for the third grade models and stochastic gradient boosting had a low R^2 for the third grade expressive model. These are adaptable and robust models that better explained the influence of lexical characteristics on children's word learning than the other models considered. Their biggest shortcoming is their "black box" nature and that they are not easily interpretable. The advanced nature of the models makes their implementation more

difficult and some level of apriori knowledge is needed, such as the proper number of trees to include or the best loss function.

Educational researchers examining word learning are interested in identifying the ways in which children learn words, the instructional programs that facilitate vocabulary acquisition, and the lexical features that may impact learning. It is imperative that researchers have results that are interpretable and actionable. Therefore, multivariate adaptive regression splines is the better model to interpret the influence of lexical characteristics on children's word learning based on the ILIAD data. It has stronger performance than linear models and shrinkage methods. SVR explains the influence comparably to MARS but it not as interpretable. Support vector regression, random forest, gradient boosting, and stochastic gradient boosting are "black box" models that are difficult to implement and interpret for researchers without expertise in them. The ensemble methods did not outperform MARS enough to justify the loss of interpretability.

MARS is a flexible and robust method of regression that can deal with some level of multicollinearity and works well with nonlinear data. It is interpretable and does not take too much prior knowledge to implement, as linear splines work for most situations. Variable selection is automatically done during model building so it can deal with a large number of parameters and continuous variable selection is performed so it is even more adaptable the discrete variable selection methods. Based on the exploration of the ILIAD data and comparison of regression models, MARS is an excellent choice for behavioral health researchers studying children's word learning. MARS has the potential to advance the field and bring new insights by giving researchers a robust, adaptable tool that can handle most types of data.

In this chapter, a strong candidate model, MARS, was proposed to analyze children's word learning. This comparison followed the same criteria for comparing models in

chapter 4 based on the assumptions, strengths, and weaknesses. A subset of the data was used to compare the models and validate the choice for selecting MARS as the strongest candidate method to regress the data.

CHAPTER SIX:

USING MARS TO PREDICT THE RELATION BETWEEN LEXICAL CHARACTERISTICS AND WORD LEARNING

Note to Reader

This chapter presents a manuscript that has been submitted to *Journal of Speech, Language, and Hearing Research* for publication and is currently under review.

Introduction

In the previous chapter we considered advanced statistical learning and machine learning models that are more robust and can improve the understanding of the ILIAD word learning data. The shrinkage methods performed well but did not stand out which may mean that the multicollinearity was not as impactful as believed or their own assumptions were too restrictive. Multivariate adaptive regression splines, support vector machines, random forest, gradient boosting, and stochastic gradient boosting all performed very well at modeling the data based on model fit metrics. Of these, MARS stands out because it is more interpretable than the other methods for regressing the data. It is not a "black box" model like random forest, gradient boosting, and stochastic gradient boosting, which do not include an underlying model that can be considered. Support vector regression is difficult to interpret because it is dependent on the support vectors and similarly does not have a tradition model. MARS does not require much a priori knowledge about the distribution of the data and works well in most situations using linear basis functions. We chose MARS to model children's word learning because it is a robust,

adaptable method for regression that has a strong balance of precision while being interpretable and relatively simple to implement.

The preceding chapters have demonstrated the merits for using MARS as an alternative to other commonly used statistical techniques to examine word learning studies. Based on the promising results of applying MARS to the ILIAD dataset, other analytic problems may be pursued this way. We have explained why MARS was a strong choice to model the data using the first-grade dataset as an example. The next step is to apply MARS to other word learning datasets to determine if it performs similarly. This chapter consists of two studies. In the first study two additional word learning datasets will be used to compare performance of different modeling techniques to validate previous findings using outcomes from a study examining Story Friends, a preschool vocabulary program, and outcomes from a study examining a kindergarten vocabulary program. In the second study, MARS will be used to predict the influence of lexical characteristics on word learning. In the preschool study, there was only one measure of word learning, a decontextualized measure. In order to compare model results between the three datasets, only the decontextualized word learning was used for analysis. The detailed model information about the most relevant characteristics will be used in a comparative analysis to explore trends found in word learning.

Outcomes that corroborate the findings in this study may have important educational implications for vocabulary instruction. If we can model similar results with word learning outcomes from studies with new participants and different words, it would strengthen our argument for using the relevant predictors to create a developmentally appropriate sequence of targets for instruction. This sequence would aid in the reduction of variability in word selection.

Study 1: Model Comparison to Validate MARS

Story Friends Preschool Data Description

Story Friends, a supplemental preschool vocabulary program (Goldstein & Kelley, 2016). Story Friends was created as a result of the ILIAD study. Many of the instructional components are similar including explicit, embedded instruction, providing multiple contexts for the words and opportunities to respond to instruction. The explicit instruction included a child-friendly definition, a contextual example of the word (related directly to the story) and decontextualized examples of the word (related to something outside of the story).

The measurement tool used to assess word learning was similar to the decontextualized measure used in the ILIAD study. Children were asked "*Tell me, what does __ mean?*" If the child did not provide a correct response a secondary contextual prompt from the story was provided (e.g., "*Ellie is enormous. Enormous means…*"). Children's responses were scored on a three-point scale, 0 for an incorrect response, 1-point for partial knowledge, and 2-points for a correct definition. Just as with the ILIAD data, the partial and full knowledge scores were collapsed into one category: learned. Thus, the outcomes for Story Friends were binary, 0-points for not learned and 1-point for learned.

The Story Friends dataset includes word learning outcomes for 72 words learned by 112 preschool children. Each vocabulary word was characterized for the following lexical characteristics: word frequency, age of acquisition, level of concreteness, neighborhood density, and phonotactic probability.

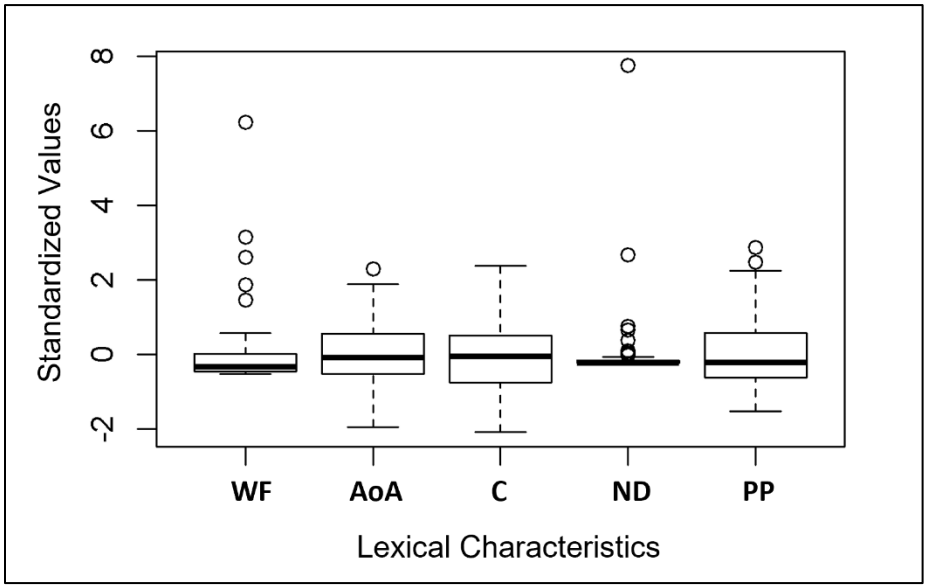
As with the ILIAD dataset, an exploratory data analysis was performed. The average percentage of preschool children who learned words was 57% (SD 19%) with a median of 56% (ranging from 19% – 89%). Descriptive statistics including mean, standard deviation, median,

range, and skew for the lexical characteristics are listed in Table 6.1. This can be visualized in Figure 6.1, the boxplot for the lexical characteristics for Story Friends words, standardized for comparison. Neighborhood density and word frequency were heavily skewed. The data was then checked for linearity, normality, multicollinearity.

Table 6.1. *Descriptive Statistics of Model Variables for Story Friends*

	M	SD	m	Min	Max	Skew
AoA	7.01	1.08	6.92	4.91	9.5	0.31
Neighborhood Density	1,424.23	6,244.15	36.84	0	4,9799.94	6.73
Concreteness	2.84	.76	2.81	1.25	4.67	0.39
Phonotactic Probability	.27	.12	.24	.08	.62	0.85
Word Freq	45.65	87.44	17.75	0.20	590.69	4.14

Figure 6.1. *Box Plot for Story Friends Model Variables (Lexical Characteristics)*



Note. WF= Word Frequency, AoA= Age of Acquisition, C= Concreteness, ND= Neighborhood Density, PP= Phonotactic Probability

The data was first checked for correlations between variables and the results can be found in Table 6.2. Word frequency and AoA are correlated according to Pearson, Kendall, and Spearman tests ($r = -.48, \tau = -.40, \rho = -.53$). Word frequency was also correlated with level of concreteness for each test ($r = -.33, \tau = -.29, \rho = -.42$) and with neighborhood density according to Kendall and Spearman ($\tau = .18, \rho = .26$). Neighborhood density is correlated with every parameter for the Kendall test and Spearman test. Overall, there is a moderate amount of correlation, which is similar to the ILIAD dataset. This strengthens the argument for continuing to use MARS over multiple and stepwise regression methods.

Tables 6.2. *Correlation between Variables for Story Friends*

Pearson	W Freq	AoA	Concrete	NDen	Phon Prob
W Freq	1				
AoA	-.48*	1			
Concrete	-.33*	.11	1		
NDen	.00	-.10	.11	1	
Phon Prob	-.05	.14	-.18	-.24*	1
Kendall					
W Freq	1				
AoA	-.40	1			
Concrete	-.29*	.09	1		
NDen	.18*	-.22*	.19*	1	
Phon Prob	.08	.13	-.13	-.41*	1
Spearman					
W Freq	1				
AoA	-.53*	1			
Concrete	-.42*	.15	1		
NDen	.26*	-.31*	.27*	1	
Phon Prob	-.11	.18	.18	-.58*	1

Variance inflation factors were calculated for the Story Friends data and can be found in Table 6.3. Using the rule of thumb that values between 1 and 5 represent moderate multicollinearity, the table shows that there exists low to moderate multicollinearity within the data. Homoscedasticity was checked using the Breusch-Pagan test and a test statistic of 5.36 was

calculated. This test statistic was not significant, therefore the data is homoscedastic. This differs from the ILIAD dataset.

Table 6.3. *Variance Inflation Factor (VIF) Test for Multicollinearity*

Variables	Tolerance	VIF
Word Frequency	.6912568	1.446640
Age of Acquisition	.7510978	1.331385
Concreteness	.8461178	1.181868
Neighborhood Density	.9334652	1.071277
Phonotactic Probability	.8968779	1.114979

Note. Lower VIF values indicate the data has lower multicollinearity.

Univariate normality was checked using the Shapiro-Wilk W test and the results can be found in Table 6.4. All of the scores were significant, meaning that none of the parameters follow a normal distribution. Looking at the W test statistics, neighborhood density and word frequency differ from normal drastically. AoA, level of concreteness, and phonotactic probability are closer to a normal distribution. These are the results that were expected based on the skew in the descriptive statistics.

Table 6.4. *Shapiro-Wilk W Test for Univariate Normal*

Parameter	W Score	*p*-value
Age of Acquisition	.97457	.1525
Neighborhood Density	.22523	< .001
Level of Concreteness	.9652	.0437
Phonotactic Probability	.93638	.0013
Word Frequency	.49676	< .001

Note. W scores can range from 0 to 1, values closer to 1 indicate data is normally distributed.

Multivariate normality was checked and the results are in Table 6.5. The Mardia skewness and kurtosis, Doornik-Hansen, Henze-Zirkler, Royston, and Energy E tests all agreed that the Story Friends data is not multivariate normal. The Story Friends data follows the ILIAD data based on the exploration of correlation, multicollinearity, univariate normality, and multivariate normality. It was found to be homoscedastic, which differs from the ILIAD first grade and full dataset which were found to be heteroscedastic. These conditions for the data justify the continued implementation of MARS for modeling the data.

Table 6.5. *Multivariate Normal Tests*

Test	Test Statistic	Multivariate Normal
Mardia Skewness	MS = 847.19*	NO
Mardia Kurtosis	MK = 32.36*	NO
Doornik-Hansen	E = 656.48*	NO
Henze-Zirkler	HZ = 3.24*	NO
Royston	HZ = 145.88*	NO
Energy	E=5.15	NO

Kindergarten Data Description

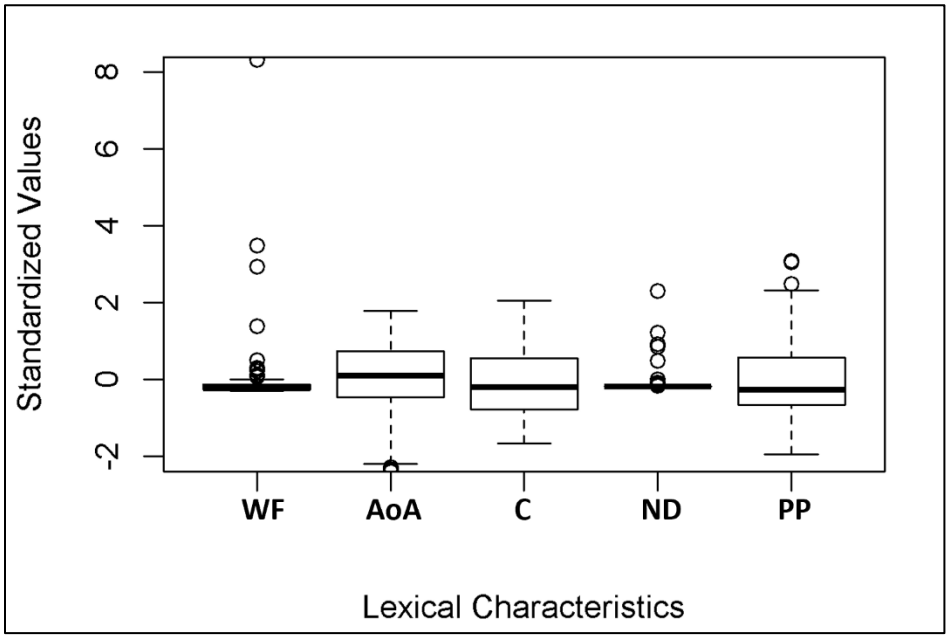
A study examining a supplemental vocabulary program was implemented with 174 kindergarteners and taught 98 vocabulary words. This program was the precursor to the vocabulary program implemented in the ILIAD study and included word learning outcomes for an expressive and decontextualized learning measures. For decontextualized learning, the average percentage of kindergarten children who learned words was 28% (SD 29%) with a median of 15% (ranging from 0% – 94%). For expressive tasks, the average percentage of kindergarten children who learned words was 23% (SD 27%) with a median of 13% (ranging from 0% – 97%). Descriptive statistics including mean, standard deviation, median, range, and skew for the lexical characteristics are listed in Table 6.6. This can be visualized in Figure 6.2,

the boxplot for the lexical characteristics for kindergarten words, standardized for comparison. Neighborhood density and word frequency were heavily skewed. The data was then checked for linearity, normality, multicollinearity.

Table 6.6. *Descriptive Statistics of Model Variables for Kindergarten*

	M	SD	m	Min	Max	Skew
AoA	8.06	2.07	8.28	3.11	11.78	-.57
Neighborhood Density	1421.19	8140.46	8.65	0	76818.95	8.26
Concreteness	3.15	.90	2.99	1.66	5	.51
Phonotactic Probability	.30	.14	.26	.02	.74	.88
Word Freq	36.61	127.99	8.82	.31	1102.98	6.61

Figure 6.2. *Box Plot for Kindergarten Model Variables (Lexical Characteristics)*



Note. WF= Word Frequency, AoA= Age of Acquisition, C= Concreteness, ND= Neighborhood Density, PP= Phonotactic Probability

The correlations between variables for the kindergarten dataset can be found in Table 6.7. Word frequency and AoA are correlated according to Pearson, Kendall, and Spearman tests $(r = -.43, \tau = -.48, \rho = -.65)$. Word frequency was also correlated with level of concreteness for each test $(r = -.31, \tau = -.23, \rho = .33)$ and with neighborhood density according to Kendall and Spearman $(\tau = .36, \rho = .52)$. Neighborhood density is correlated with every parameter for the Kendall test and Spearman test. Overall, there is a moderate amount of correlation, which agrees well with the ILIAD and Story Friends datasets. This strengthens the argument for continuing to use MARS over multiple and stepwise regression methods.

Tables 6.7. *Correlation between Variables for Kindergarten*

Pearson	W Freq	AoA	Concrete	NDen	Phon Prob
W Freq	1				
AoA	-.43*	1			
Concrete	.31*	-.62*	1		
NDen	.03	-.29*	.27*	1	
Phon Prob	-.18	.36*	-.25*	-.24*	1
Kendall					
W Freq	1				
AoA	-.48*	1			
Concrete	.23*	-.33*	1		
NDen	.36*	-.34*	.26*	1	
Phon Prob	.20*	.26*	-.15	-.23*	1
Spearman					
W Freq	1				
AoA	-.65*	1			
Concrete	.33*	-.47*	1		
NDen	.52*	-.49*	.37*	1	
Phon Prob	-.31*	.39*	-.21*	-.33*	1

Variance inflation factors for the kindergarten data can be found in Table 6.8. The table shows that there exists low to moderate multicollinearity within the data. Homoscedasticity was checked using the Breusch-Pagan test resulting in a test statistic of 4.39 for decontextualized

learning and 17.2 for expressive tasks. The decontextualized test statistics was not significant, meaning the data is homoscedastic.

Table 6.8. *Variance Inflation Factor (VIF) Test for Multicollinearity (Kindergarten)*

Variables	Tolerance	VIF
Word Frequency	.7980841	1.253001
Age of Acquisition	.5040246	1.984030
Concreteness	.5970255	1.674970
Neighborhood Density	.8693382	1.150300
Phonotactic Probability	.8492778	1.177471

Note. Lower VIF values indicate the data has lower multicollinearity.

Univariate normality was checked using the Shapiro-Wilk W test and the results can be found in Table 6.9. None of the parameters follow a normal distribution based on the significance for each test. Looking at the W test statistics, neighborhood density and word frequency differ from normal drastically, while AoA, level of concreteness, and phonotactic probability are closer to a normal distribution. These are the results that were expected based on the skew in the descriptive statistics.

Table 6.9. *Shapiro-Wilk W Test for Univariate Normal*

Parameter	W Score	*p*-value
Age of Acquisition	.95963	.0043
Neighborhood Density	.16607	< .001
Level of Concreteness	.93852	< .001
Phonotactic Probability	.94674	< .001
Word Frequency	.26036	< .001

Note. W scores can range from 0 to 1, values closer to 1 indicate data is normally distributed.

Multivariate normality was checked, and the results are listed in Table 6.10. The Mardia skewness and kurtosis, Doornik-Hansen, Henze-Zirkler, Royston, and Energy E tests all agreed that the kindergarten data is not multivariate normal. The kindergarten data is very similar to the

Story Friends and ILIAD dataset based on the exploration of correlation, multicollinearity, univariate normality, and multivariate normality. It was found to be homoscedastic, which differs from the ILIAD first grade and full dataset which were found to be heteroscedastic, but agrees with the second grade data and Story Friends data. These conditions for the data justify the continued implementation of MARS for modeling the data.

Table 6.10. *Multivariate Normal Tests (Kindergarten)*

Test	Test Statistic	Multivariate Normal
Mardia Skewness	MS = 2001.19*	NO
Mardia Kurtosis	MK = 68.28*	NO
Doornik-Hansen	E = 872.70*	NO
Henze-Zirkler	HZ = 5.30*	NO
Royston	HZ = 192.51*	NO
Energy	E=8.52*	NO

Results

Two new datasets were modeled to compare the performance and validate the decision to choose MARS for modeling children's word learning. Each technique was used to model the Story Friends preschool data and the kindergarten dataset without expressive tasks and decontextualized learning outcomes. Each model was compared using the coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). The results for this comparison can be found in Table 6.11. As with the ILIAD data comparison, MARS better describes the data than multivariate linear regression, stepwise regression, and the shrinkage methods. Regression trees and support vector regression had similar performance to MARS for each of the datasets. Random forests did not explain as much of the variance as MARS for each dataset but had smaller error measures. Gradient boosting machines slightly outperformed MARS for each dataset and stochastic gradient boosting machines outperformed MARS for the Story Friends data but were similar for the kindergarten data.

Discussion

The results from the comparison are similar to the performance for each model using the ILIAD dataset. MARS outperforms multivariate linear regression and stepwise regression based on every fit metric. Based on the exploration of the data, this was expected because many of assumptions they rely on are not true for the data. Similarly, the shrinkage methods had weaker results than MARS. Support vector regression was originally chosen because it has comparable adaptability and robustness to MARS. This is supported by the results of the comparison, performing equally well to MARS for every dataset. While support vector regression would be a strong choice for modeling children's word learning, the lack of an easy to interpret model and the complexity of choosing hyperparameters makes MARS strong alternative.

Regression trees performed well and the resulting tree diagrams give a novel interpretation of the data. The nature of tree diagrams leaves many gaps in the outputs for prediction and regression trees do not have an easily interpretable model. Because of this, analysis of lexical characteristics makes discrete jumps and some information for decision making may be lost. The more advanced tree based methods all outperformed MARS to some degree. As with the ILIAD comparison, their "black box" nature means the results will be less interpretable for researchers. Random forest and stochastic gradient boosting machines often require larger datasets, so their performance may be more variable.

Multivariate adaptive regression splines continues to be a strong choice for modeling children's word learning. It is adaptable and robust, balancing performance with interpretability. While some models may have less error or explain more of the variance, the results may not be as actionable as MARS and simpler models. The consistent performance of MARS validates it as the best model for children's word learning research.

Table 6.11. *Model Comparison for Kindergarten and Story Friends*

	MARS	MR	SR	Ridge	LASSO	PLS	PCR	SVR	CART	RF	GBM	SGBM
Story Friends												
R^2	.30	.14	.09	.12	.09	.14	.14	.31	.30	.12	.71	.88
MAE	.13	.15	.16	.15	.16	.15	.15	.12	.13	.07	.08	.05
MSE	.03	.03	.03	.03	.03	.03	.03	.03	.03	.01	.01	.00
RMSE	.16	.18	.18	.18	.18	.18	.18	.16	.16	.09	.10	.07
Kindergarten Decontextualized												
R^2	.84	.71	.71	.68	.70	.71	.66	.81	.76	.75	.91	.85
MAE	.09	.12	.12	.13	.12	.12	.13	.09	.10	.05	.06	.08
MSE	.01	.02	.02	.03	.02	.02	.03	.02	.02	.00	.01	.01
RMSE	.11	.15	.15	.16	.16	.15	.17	.12	.14	.06	.09	.11
Kindergarten Expressive												
R^2	.81	.63	.63	.63	.62	.63	.63	.78	.64	.71	.83	.83
MAE	.09	.12	.12	.12	.12	.12	.12	.08	.08	.05	.08	.08
MSE	.01	.03	.13	.03	.03	.03	.03	.02	.01	.00	.01	.01
RMSE	.11	.16	.16	.16	.16	.16	.16	.13	.11	.07	.11	.11

Note. MARS= Multivariate Adaptive Regression Splines, MR= multivariate linear regression, SR= Stepwise Regression, LASSO= Least Absolute Shrinkage & Selection Operator, PLS= Partial Least Squares, PCR= Principal Component Regression, SVR= Support Vector Regression, CART= Classification & Regression Tree, RF= Random Forest, GBM= Gradient Boosting Machines, SGBM= Stochastic Gradient Boosting Machines, MAE= Mean Absolute Error, MSE= Mean Square Error, RMSE= Root Mean Square Error.

Study 2: Using Multivariate Adaptive Regression Splines (MARS) to Examine the Influence of Lexical Characteristics on Word Learning

The aim for study 2 was to use MARS to model the influence of lexical characteristics on children's decontextualized word learning from the ILIAD, Story Friends, and kindergarten datasets.

ILIAD Results

Table 6.12 lists the descriptive statistics for each grade level. For each of the grade-level models, the following results are presented: variable importance and selection criteria, the final model including regressions and associated hinges for the statistically significant lexical characteristics related to word learning, and the graphs associated with each model. Each graph represents the effect of each variable on word learning while all others were held constant. The graphs are ordered by calculated importance.

First Grade

For the first-grade model, the percentage of children who learned the target vocabulary words (n= 143 words) taught in first grade were entered into the model. The resulting model identified five lexical characteristics as relevant predictors of word learning listed in Table 6.13 by order of importance. The most important variable was age of acquisition followed by level of concreteness, neighborhood density, word frequency, and finally phonotactic probability. The final model included seven basis functions listed in Table 6.14.

Using the graphical representation of model results in Figure 6.3, we can examine the relation between word learning and each characteristic while all others are held constant. Age of acquisition remained steady until the hinge at 5.37 years old where the percentage of children who learned words decreased rapidly until the age of 7.81. There was a slight jump in learning between age of acquisition ratings of 7.81 and 8.45 years old, and then slowly decreased as age of acquisition goes up. The percent of children who learned words remained steady for words with concreteness ratings from 1.5 to 3, and then slowly increased after the hinge at 3. Words that were more concrete were learned by more first graders. Trends for neighborhood density,

word frequency, and phonotactic probability remained neutral indicating these variables did not seem to impact the number of words learned by children. The variability in learning was mostly accounted for by age of acquisition and level of concreteness.

Table 6.12. *Descriptive Statistics for ILIAD Model Variables (Lexical Characteristics)*

First Grade (n= 143)	M	SD	m	Min	Max	Skew
Word Learning	26%	29%	14%	1%	99%	–
AoA	8.80	2.17	9.06	3.25	13.61	-.46
N_Den	1845.32	8474.12	6.49	0	69210.62	6.23
Conc_Mean	2.96	.97	2.76	1.50	5	.64
Phon_Prob	.22	.12	.21	.03	.52	.57
SUBTLwf	19.29	57.94	6.90	.27	509.37	6.67
Second Grade (n= 126)						
Word Learning	38%	28%	27%	3%	97%	–
AoA	8.63	2.29	8.63	3	13.41	-.23
N_Den	1106.11	5560.46	8.53	0	45721.92	6.92
Conc_Mean	2.89	1.00	2.63	1.46	4.97	.72
Phon_Prob	.24	.15	.21	.02	.66	.76
SUBTLwf	31.37	105.91	7.63	.02	801.82	6.07
Third Grade (n= 108)						
Word Learning	22%	15%	18%	3%	74%	–
AoA	10.30	1.41	10.25	6.75	14.5	.15
N_Den	771.50	6666.04	1.41	0	69210.62	10.02
Conc_Mean	2.39	.62	2.29	1.43	4.15	.81
Phon_Prob	.24	.14	.21	.03	.72	1.37
SUBTLwf	4.82	6.39	2.46	.08	35.65	2.57

Note: M= mean, SD= standard deviation, m= median, min= minimum value, max= maximum value, n= number of words, AoA= age of acquisition, N_Den= neighborhood density, Conc_Mean= level of concreteness, Phon_Prob=phonotactic probability, SUBTLwf= word frequency. Skew is not reported for word learning.

Table 6.13. *Importance of Explanatory Variables in the First Grade MARS Model*

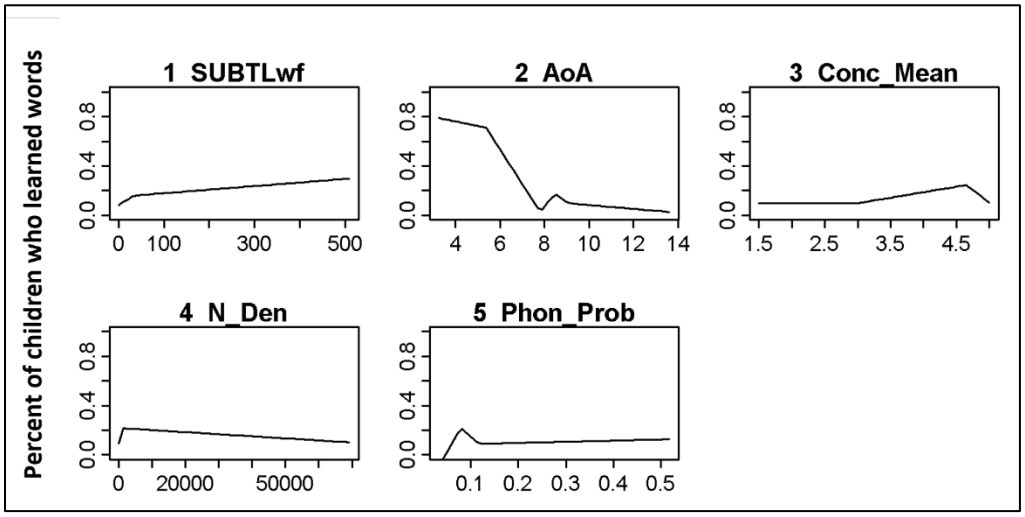
Variable	nsubsets	GCV	RSS
Age of Acquisition	7	100	100
Level of Concreteness	5	22.1	26
Neighborhood Density	4	16	20.3
Word Frequency	2	8.1	12.1
Phonotactic Probability	1	5.5	8.3

Table 6.14. *MARS Results for First Grade Decontextualized Word Learning*

Predictor	Type	Hinge Location	Coefficient
(Intercept)			0.91
AoA	Right	5.37	-0.27
AoA	Right	7.81	0.42
AoA	Right	8.45	-0.17
Concrete	Right	3.00	0.07
N Den	Left	126.04	-0.001
Word Freq	Left	32.22	-0.003
Phono Prob	Left	0.08	-3.05

Note. Word Freq= Word Frequency; AoA= Age of Acquisition; Concrete= Level of Concreteness; N Den= Neighborhood Density; Phono Prob= Phonotactic Probability`

Figure 6.3. *Variable Plot for First Grade Decontextualized Learning using MARS*



Note. Scale for each x-axis differs based on lexical characteristic values.

Second Grade

For the second-grade model, the percentage of children who learned the target vocabulary words (n= 126 words) taught in second grade were entered into the model. In Table 6.15, the most important variable was age of acquisition followed by level of concreteness, word frequency, neighborhood density, and finally phonotactic probability. The final model included nine basis functions listed in Table 6.16 and is depicted in Figure 6.4. Based on age of acquisition, the percentage of children who learned words decreased slowly until age 9.35 where

learning seemed to remain neutral until 11.44. Learning began to increase for words with age of acquisition ratings older than 11.44. Learning steadily increased as words became more concrete (values closer to 5). Word frequency remained neutral; learning did not seem to vary for words as frequency rates increased. The percent of children who learned words steadily declined as neighborhood density values increased, that is why there is no hinge present in the figure. Words in denser neighborhoods were more difficult for children to learn compared to words in sparser neighborhoods. Phonotactic probability had a slightly varied impact on the percent of children who learned words; learning drops rapidly as probabilities increased to .07 and then remained mostly neutral with minimal increases and decreases between hinges at probabilities .22 and .45.

Table 6.15. *Importance of Explanatory Variables in the Second Grade MARS Model*

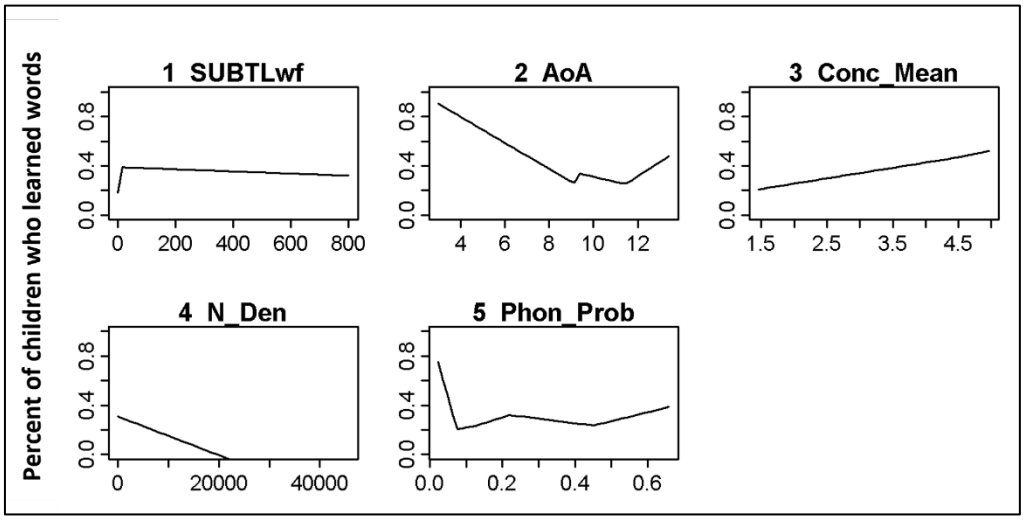
Variable	nsubsets	GCV	RSS
Age of Acquisition	9	100	100
Level of Concreteness	8	35.8	44.8
Word Frequency	7	26.8	37.2
Neighborhood Density	6	16.3	29.6
Phonotactic Probability	5	16.9	27.8

Table 6.16. *MARS Results for Second Grade Decontextualized Word Learning*

Predictor	Type	Hinge Location	Coefficient
(Intercept)			-3.62
AoA	Left	9.35	.09
AoA	Right	11.44	.10
Concrete	Left	4.44	-.09
Word Freq	Left	12.35	-.02
N Den	Right		-.00001
Phono Prob	Right	.07	11.60
Phono Prob	Right	.22	-1.41
Phono Prob	Left	.45	10.62
Phono Prob	Right	.45	-9.48

Note. Word Freq= Word Frequency; AoA= Age of Acquisition; Concrete= Level of Concreteness; N Den= Neighborhood Density; Phono Prob= Phonotactic Probability`

Figure 6.4. *Variable Plot for Second Grade Decontextualized Learning using MARS*



Note. Scale for each x-axis differs based on lexical characteristic values.

Third Grade

For the third-grade model, word learning data for 108 words taught in third grade and the lexical characteristics describing those words were entered into the model. MARS identified age of acquisition as the most important variable, followed by neighborhood density, word frequency, and finally level of concreteness, listed in Table 6.17. Results of the third-grade model are listed in Table 6.18 and depicted in Figure 6.5. The percentage of children who learned words remained constant until an AoA rating of 9.67, then slowly decreased as AoA increased. Learning steadily increased as neighborhood density values increase, as words in denser neighborhoods were easier for children to learn compared to words in sparser neighborhoods. As the word frequency measures increased, so do the percentage of children who learned words. Words that were more abstract were slightly more difficult for children to learn, but once words reached a concreteness rating of 2.3 learning remained neutral. It appears that level of concreteness may have had a small impact on third grade students' word learning.

Table 6.17. *Importance of Explanatory Variables in the Third Grade MARS Model*

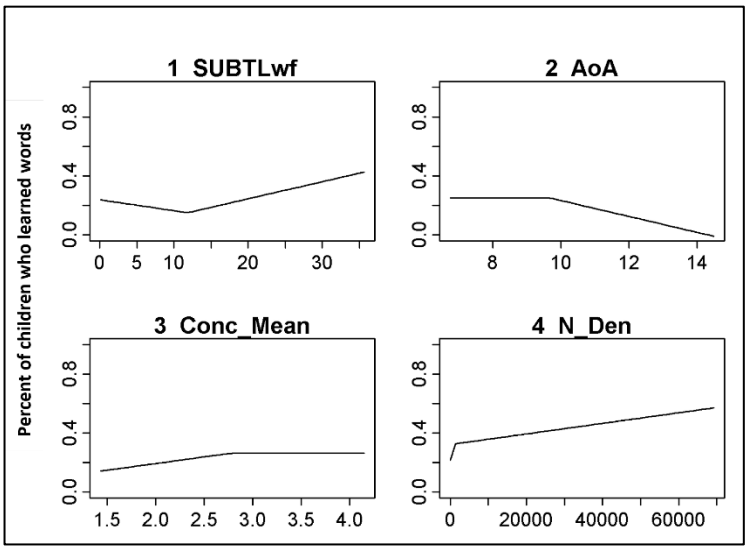
Variable	nsubsets	GCV	RSS
Age of Acquisition	7	100	100
Neighborhood Density	5	62.4	68.8
Word Frequency	4	55.5	61.8
Level of Concreteness	4	53	59.3

Table 6.18. *MARS Results for Third Grade Decontextualized Word Learning*

Predictor	Type	Hinge Location	Coefficient
(Intercept)			.12
AoA	Right	9.67	-.05
N Den	Right		.001
N Den	Right	82.79	-.001
Word Freq	Left	11.84	.007
Word Freq	Right	11.84	.01
Concrete	Right	2	.31
Concrete	Right	2.3	-.30

Note. Word Freq= Word Frequency; AoA= Age of Acquisition; Concrete= Level of Concreteness; N Den= Neighborhood Density; Phono Prob= Phonotactic Probability`

Figure 6.5. *Variable Plot for Third Grade Decontextualized Learning using MARS*



Note. Scale for each x-axis differs based on lexical characteristic values.

Story Friends Preschool Results

The percentage of children who learned each target word was entered into the model. Through variable selection, the most important characteristics related to preschoolers' word learning were level of concreteness, age of acquisition, and word frequency, listed in Table 6.19. Age of acquisition and word frequency were both 71% - 75.4% as important relative to level of concreteness. Neighborhood density and phonotactic probability were not found to be important and thus not included in the model.

Table 6.19. *MARS Variable Selection for Story Friends*

Variable	nsubsets	GCV	RSS
Level of Concreteness	5	100	100
Age of Acquisition	3	71	75.4
Word Frequency	3	71	75.4

The final model includes six basis functions reported in Table 6.20. Level of concreteness was the most related to word learning. Words that were more concrete were easier for children to learn than words that were more abstract. Up until the hinge at 2.43, level of concreteness did not impact learning and then an interesting artifact can be seen where learning changes rather quickly for words with concreteness levels between 2.6 and 2.9. Learning then slowly increased as words became more concrete (concreteness ratings > 2.9). Age of acquisition was negatively related to preschoolers' word learning. There was a steady decline in learning as age of acquisition ratings got older until the hinge at approximately 8 years, where AoA did not seem to further impact learning. Word frequency also had a negative impact on word learning; as frequency ratings increased, learning decreased until the hinge at 81.03. Frequency measures

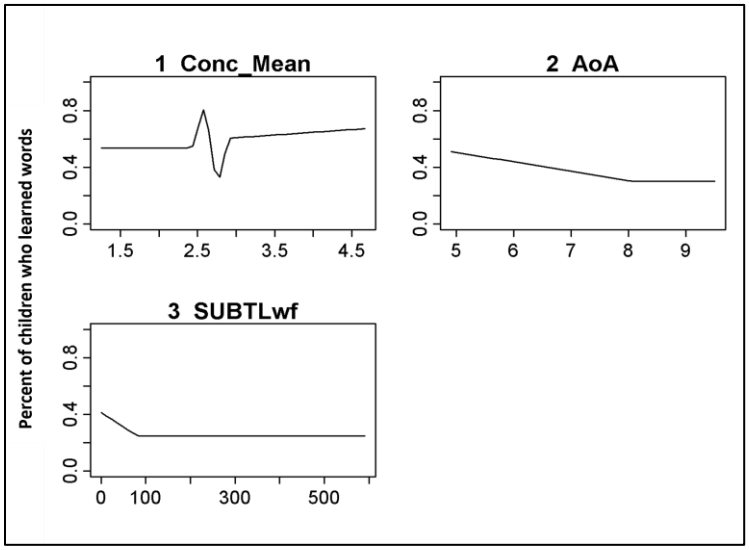
greater than 81 did not seem to impact word learning. The partial dependency plot for the variables can be found in Figure 6.6, where all other variables are held constant.

Table 6.20. *MARS Results for Story Friends*

Predictor	Type	Hinge Location	Coefficient
(Intercept)			.33
Concrete	Right	2.43	1.83
Concrete	Right	2.6	-5.84
Concrete	Right	2.75	6.41
Concrete	Right	2.9	-2.36
AoA	Left	8.05	.07
Word Freq	Left	81.03	.002

Note. Word Freq= Word Frequency; AoA= Age of Acquisition; Concrete= Level of Concreteness; N Den= Neighborhood Density; Phono Prob= Phonotactic Probability`

Figure 6.6. *Variable Plot for Story Friends Decontextualized Learning using MARS*



Note. Scale for each x-axis differs based on lexical characteristic

Kindergarten Results

The percentage of children who learned each target word was entered into the model. Through variable selection, the most important characteristics related to kindergarteners' word learning were age of acquisition, level of concreteness, and word frequency, listed in Table 6.21.

Neighborhood density and phonotactic probability were not found to be important and thus not included in the model.

Table 6.21. *MARS Variable Selection for Kindergarten*

Variable	nsubsets	GCV	RSS
Age of Acquisition	7	100	100
Level of Concreteness	6	20.5	28.3
Word Frequency	3	9.1	16.8

The final model includes seven basis functions reported in Table 6.22. Age of acquisition was the most related to word learning. Words that were rated as younger were easier for children to learn than words that were rated as learned later. Up until the hinge at 5.37, age of acquisition did not impact learning and then rapidly declines until the hinge at 6.65 where a short increase in learning occurs until the hinge at 7.58 where learning again slowly declines. Level of concreteness had a neutral impact on word learning until the hinge at 4.31 where learning rapidly increased. Words that were more concrete were easier for kindergarteners to learn. Word frequency did not seem to impact word learning after the hinge at 2.59. The partial dependency plot for the variables can be found in Figure 6.7, where all other variables are held constant.

Table 6.22. *MARS Results for Kindergarten*

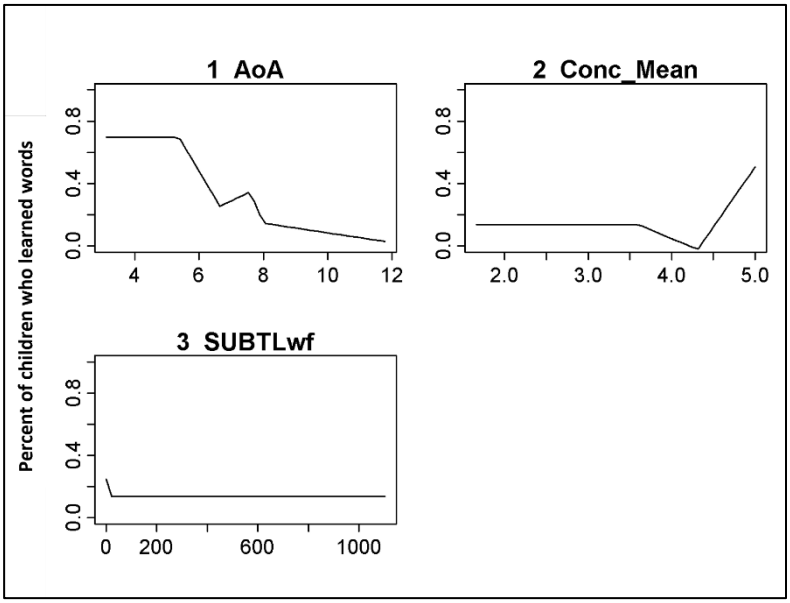
Predictor	Type	Hinge Location	Coefficient
(Intercept)			.70
AoA	Right	5.37	-.35
AoA	Right	6.65	.44
AoA	Right	7.58	-.58
AoA	Right	8	.45
Concrete	Right	3.61	-.23
Concrete	Right	4.31	1.00
Word Freq	Left	2.59	.05

Note. Word Freq= Word Frequency; AoA= Age of Acquisition; Concrete= Level of Concreteness; N Den= Neighborhood Density; Phono Prob= Phonotactic Probability

Variable Importance

Table 6.23 shows what variables were most important for each MARS model. The preschool model found level of concreteness to be the most important variable, while the kindergarten, first-, second-, and third-grade models all determined the most important lexical characteristic was age of acquisition. For three out of the four models, level of concreteness was the second most important variable. Interestingly, the third-grade model indicated neighborhood density as the second most important variable. Phonotactic probability was included in the first and second grade models, but it had almost no impact on learning.

Figure 6.7. *Variable Plot for Kindergarten Decontextualized Learning using MARS*



Note. Scale for each x-axis differs based on lexical characteristic

Goodness of Fit

The outcome of the MARS models can be critiqued using goodness of fit measures listed in Table 6.24. Based on the results of the error analyses the grade level models performed well according to the error measures. For the first and second grade models the variables did very well explaining the variance in word learning ($R^2_{1st} = 0.84, R^2_{2nd} = 0.72$) but the third-grade model did not ($R^2_{3rd} = 0.42$).

Table 6.23. *Variable Importance Across Grade Levels*

Importance	Preschool	Kindergarten	1st Grade	2nd Grade	3rd Grade
1	Level of Concretenes s	Age of Acquisition	Age of Acquisition	Age of Acquisition	Age of Acquisition
2	Age of Acquisition	Level of Concreteness	Level of Concreteness	Level of Concreteness	Neighborhood Density
3	Word Frequency	Word Frequency	Neighborhood Density	Neighborhood Density	Word Frequency
4			Word Frequency	Word Frequency	Level of Concreteness
5			Phonotactic Probability	Phonotactic Probability	

Table. 6.24. *Goodness of Fit Results*

Error Metric	Preschool	Kindergarten	1st Grade	2nd Grade	3rd Grade
R^2	0.30	0.84	0.84	0.72	0.42
Mean Absolute Error (MAE)	0.13	0.12	0.09	0.12	0.08
Mean Standard Error (MSE)	0.03	0.02	0.01	0.02	0.01
RMSE	0.16	0.15	0.12	0.15	0.11
General Cross Validation (GCV)	0.04	0.02	0.01	0.03	0.01

Discussion

Lexical Characteristics & Word Learning

A secondary data analysis of three investigations examining the effects of supplemental vocabulary interventions were conducted using MARS to identify the relations between lexical characteristics of vocabulary words and the word learning outcomes from preschool, kindergarten, first, second, and third grade students. The lexical characteristics examined were age of acquisition, neighborhood density, level of concreteness, phonotactic probability, and word frequency.

Age of Acquisition. The MARS analysis revealed significant relations between age of acquisition and students' vocabulary learning in the preschool, kindergarten, first-, second-, and third- grade models. We found that words with a younger age of acquisition rating were easier for children to learn than words with older age of acquisition ratings. Our findings support the results of a lexical access study by Newman and German (2002) who found children had an easier time naming words with lower age of acquisition. Although this seems rather intuitive, and

somewhat circular, this is an interesting factor to discuss. The level of importance attributed to age of acquisition is surprising considering the nature in which these ratings were obtained. Adults were asked to recall the age at which they learned a word. *Learned* was defined as understanding the word if others used it, but that they did not necessarily use it themselves. This can be a difficult task, especially when trying to recall learning at a very young age. Yet researcher have examined the validity of this and found that adult ratings of age of acquisition are valid (Gilhooly & Gilhooly, 1980; Gilhooly & Logie, 1980). Findings from this study reinforce age of acquisition as a reliable metric.

In second grade there was an increase in learning at the 11.44 hinge. This increase is an interesting artifact. It could be that the word(s) had other contributing factors, like a higher level of concreteness, that lead to increased learning. The definition and/or the contexts used for instruction may have also contributed to the increase in learning. As children progressed through grade levels (got older) the AoA at which learning began to decrease seemed to progress as well, from AoA ratings of 5 years-old in preschool up to almost 10 years-old in third grade.

Now that we know age of acquisition was strongly related to sophisticated vocabulary learning of children from a range of grade levels, additional analyses and studies are warranted to discover the range of AoA ratings that lead to optimized learning for each grade level. Because of the way MARS models data, we have detailed information about how age of acquisition impacts learning using hinges. When designing future studies, the hinge data could help when selecting words for instruction by pinpointing the exact age range most appropriate for each grade level. This selection of words would be more precise than using general linear trends. It may be that teachers should focus instruction on words acquired later (within reason given the age of students) because they are more difficult for children to learn than words that are acquired at an earlier age.

Level of Concreteness. MARS modeled level of concreteness as the second most important lexical characteristic related to word learning in the kindergarten, first and second grade-level models. Our results indicate, that for children in preschool, kindergarten, first, and

second grades, words that were more concrete, or high in imageability, were easier to learn than words that were more abstract, meaning they were more difficult to explain and picture. The third-grade model selected concreteness as the least important variable, it did not seem to significantly impact word learning. Interestingly, MARS modelled level of concreteness as the most important variable related to preschoolers' word learning. Descriptive statistics were examined to determine if the average concreteness level of the words taught in preschool, first, second, and third grades differed significantly. If there were differences, it could explain the differences in the model's selection of important variables. The average concreteness levels did not differ greatly across grade levels, so there may be something innately different about the age of children, how they acquire new vocabulary terms, and what lexical characteristics influence learning the most.

The hinge data provided insightful information about the underlying process of word learning in regard to abstract and concrete concepts taught in preschool, kindergarten, first, and second grades. Our findings are supported by prior research that found imageability predicted preschoolers' word learning (Hadley et al., 2021). Also, more imageable words were learned earlier and more easily than words that were less imageable (McDonough et al., 2011). Again, this finding is rather intuitive. Words that are more concrete have specific meanings, whereas words that are more abstract often have nuanced meanings that depend on context. Children can acquire more abstract terms, but if they have no referent to associate the word with, it can be difficult to retain the word's meaning. It could be that as children age, their life experiences make them well-suited to understand and describe more abstract concepts. This could explain why concreteness did not impact word learning in third grade. Further research is warranted to explore this phenomenon to better understand word learning across a larger group of children to determine if there are underlying processes that facilitate the acquisition of abstract terms. This characteristic coupled with age of acquisition could facilitate the creation of a developmental sequence for vocabulary instruction.

Word Frequency. Word frequency was included as an important variable in all of the grade models. However, when examining level of variable importance, word frequency was 71% to 75% as important to preschoolers' word learning as level of concreteness but had little impact on kindergarten word learning (9% - 17% relative importance) and first grade word learning (8% - 12% relative importance). In the other grade levels, word frequency's level of importance increased from 26% in second grade to 55% in third grade. For preschool, kindergarten, first, and second grades, it appeared that as word frequency increased, learning mostly remained the same. It is important to remember that word frequency, when combined with other more important variables like age of acquisition and level of concreteness, the majority of the variability in learning was accounted for by these variables and less by the frequency of a word.

The words in this study did not include words with very high measures of word frequency, so our findings must be interpreted carefully due to the restricted range of frequencies. In all three studies, words were selected using Beck and colleagues' (2002) framework for word selection. They recommend choosing target vocabulary words children will not likely hear in everyday conversation, but ones that would have high utility and appear later in academic texts. Other researchers have found that words that occur more frequently were easier for children to name in a lexical access study (Newman & German, 2002).

Word frequency values ranged from .02 to 801 and were heavily skewed for all grade levels. While the words chosen may not seem to have a lower frequency among adults, they may have infrequent use by young children. Further analyses should investigate word frequency norms for children by examining childhood literature or television shows and movies made for children. Either of these methods would mirror popular adult word frequency norms derived from print or television and movies (Brysbaert & New, 2009; Francis & Kučera, 1982). If differences existed between the frequency norms of children and adults, it would allow for a more robust measure used to examine the relations between frequency and young children's vocabulary learning.

Phonotactic Probability & Neighborhood Density. MARS included neighborhood density and phonotactic probability in the first- and second-grade models. Both variables were included in the first and second grade models, and only neighborhood density was included in the third-grade model. Neither variable was included in the preschool or kindergarten models. In the first and second grade models, just like word frequency, it could be that when grouped with other more influential lexical characteristics, such as age of acquisition and level of concreteness, the impact of neighborhood density and phonotactic probability on word learning was minimal. Interestingly, in the third-grade model, neighborhood density was the second most important variable related to word learning. There was a sharp learning increase until the hinge at 82.79, and then a steady increase in learning as neighborhood density measures increased.

Previous research has found a relation between phonotactic probability and neighborhood density (Hoover et al., 2010). However, results of this analysis did not reveal a relationship between the two. Additionally, there was little-to-no relation among these factors and word learning in the preschool, kindergarten, first- and second- grade models. These findings are similar to that of Storkel and colleagues (2006) who were unable to demonstrate an interaction between phonotactic probability and neighborhood density in a study examining adult word learning. Our findings could be attributed to the correlation between word length and these lexical characteristics, since most of the words in our analysis varied in length and were multisyllabic. When words vary in length, problems in analysis and interpretation can occur (Storkel, 2004). This could explain why we did not find significant relations between word learning and phonotactic probability and neighborhood density.

Phonotactic probability is directly affected by word length. It is calculated using the sum of log values, which is equivalent to the log of the values multiplied. When multiplied together, values in this range will always decrease. This leads to a decrease in phonotactic probability as word length increases. Word length was not a factor controlled for in this study. Phonotactic probabilities for the words in our analysis ranged from 0.01-0.08. These small probabilities were not significantly related to word learning nor to neighborhood density.

Neighborhood density is negatively correlated to word length. The density increases for shorter words that have more similar neighbors, and decreases in density as word length increases, where longer words have fewer similar neighbors. Because our words varied in length, we had a large range of density measures, from 0 to 69,210.62. About half of the words (53%) had a neighborhood density of 0-5, and only 17% of the words had density measures over 100. In practice, when vocabulary targets are chosen for instruction, words taught at the same time should be semantically and phonetically distinct from one another to avoid confusion. Because of this, phonotactic probability and neighborhood density are unlikely to be important factors when selecting sophisticated vocabulary targets.

Multivariate Adaptive Regression Splines (MARS)

Although MARS is not a popular method for analysis in educational research, it is a better option compared to other well-established analyses. This is especially important when considering the nature of the data being analyzed. Data that are nonlinear, have multicollinearity, mixed variables, and other factors like lack of homoscedasticity can be a problem for most simpler methods. Student data tend to be messy, and while linear models will give results, they may be less reliable.

Given the ILIAD, Story Friends, and kindergarten datasets, MARS was a better model compared to other linear methods. The ILIAD data were shown to be piecewise and multivariate non-normal, skewed, and had some multicollinearity. Likewise, the Story Friends and kindergarten data do not follow a normal distribution, is skewed, and has multicollinearity. The weaknesses of most approaches are the assumptions they depend on to work properly. MARS does not rely on base assumptions, which is one of its strengths. Other complex methods may perform equally well but MARS is beneficial for its interpretable nature. Alternatives such as ensemble methods (e.g., random forest regression and gradient boosting machines) may be more accurate but are considered a “black box” approach. These provide outcomes, but not insight into how the decisions were made. Another major consideration is computational complexity for modeling the data. MARS is a robust and adaptable method for modeling that also very efficient.

The ILIAD, Story Friends, and kindergarten data are an important example of why it is imperative to check for non-linearity in data before deciding on a method for analysis. Many tools are tried and true but may not be as reliable based on their starting assumptions, specifically the assumption that all data are normally distributed and linearly dependent. Similarly, the metrics used to consider how well a model fits the data are reliant on certain assumptions, like with R^2. With this in mind, MARS is a strong alternative to linear methods for analyzing word learning data and minimizing error was the appropriate goodness of fit metric to test the results.

In this chapter, our contribution to children's word learning was analyzing the influence of lexical characteristics on children's word learning using MARS. This expands on prior research on the subject by more precisely modeling the impact for each lexical characteristic. This allows us to more accurately explain the effect each lexical characteristic has on children's word learning, as well as finding interesting artifacts in the word selection missed by ordinary linear regressions. To aid researchers with implementing MARS in future research, a step-by-step guide was included in the Appendix.

CHAPTER SEVEN:

CONCLUDING DISCUSSION

Statistical rigor and collaboration are lacking in many fields of research. The collaboration of interdisciplinary fields with statisticians can help limit poor methodological and statistical practices (Sainani et al, 2021; Veldkamp et al, 2014). While guidelines for statistical consulting are lacking and there are many challenges (Khamis & Mann, 1994), it is vital to work with researchers to increase statistical training. While there exists a push for cross collaboration, certain fields of research are often not considered and left out. Early education intervention is one such field that is overlooked and therefore lacks the use of advanced statistical techniques (Snyder et al., 2002). This work is result of such collaboration.

While advanced statistical techniques have begun to show up in educational research, educational intervention research focusing on vocabulary acquisition has remained reliant on simpler methods to analyze word learning. These methods, such as multivariate linear regression and stepwise regression, are easy to implement and interpret, but make many assumptions about the data. These assumptions can lead to unreliable results. Their simplicity can overlook finer details that are missing from the final model, making it challenging to interpret results. This does not mean these results are not relevant, but alternative methods may provide additional information about relations modeled.

The purpose of this dissertation was to examine statistical and machine learning methods that deal with the types of data that may be encountered during word learning research. Word

learning data can be “messy,” and issues such as skew, multicollinearity, and non-normal distributions can impact the ways in which models are selected. An effort was made to focus on aspects of data that may cause simpler models to be unreliable. Shrinkage methods were first examined to deal with multicollinearity and variable selection, while not deviating far from familiar concepts like multivariate linear regression. These methods are not as predictive compared to other advanced techniques and did as well as multivariate linear regression. This means that the multicollinearity is likely not the cause for poor model main effects, but that the skew of the data is.

More advanced models such as support vector regression, random forest, and gradient boosting machines were examined to handle aspects like skew and the non-normal nature of the data. While these models performed well, they have their own shortcomings. They are harder to interpret and understand and requires a priori knowledge and expertise when making decisions about modeling, like implementing hyperparameter tuning. Compared to these other methods, multivariate adaptive regression splines are a strong balance of model performance and interpretability for most “real world” data. It is not reliant on any starting assumptions and can be adapted with minimal prior knowledge. MARS uses variable selection when building models. This creates simpler, more interpretable models. MARS is computationally efficient compared to most other methods, so it can be applied to larger datasets without becoming an unreasonably time-consuming endeavor.

Results of this dissertation indicate age of acquisition and level of concreteness were the most influential factors related to word learning. These results were consistent across several grades, from preschool to third grade. MARS provided hinge data that indicated changes in word learning relevant to lexical characteristic values. Age of acquisition confirmed that children

learned words at their age and then dropped rapidly as AoA ratings increased. Words that were more abstract were harder for children to learn, but as children got older, they were able to learn more abstract words.

By implementing sophisticated analyses, results have the potential to elucidate additional relations among lexical characteristics and word learning to strengthen outcomes from vocabulary instruction and intervention studies. The outcomes from this dissertation have the potential to identify the lexical characteristics that contribute to the overall likelihood vocabulary words will be learned by children. Using lexical factors, we can create a systematic approach to word selection. By utilizing these lexical characteristics for word organization, it is possible to create a more unified, developmentally appropriate sequence of vocabulary targets used for instruction. This will improve vocabulary learning and has the potential to close the achievement gap among from vulnerable populations.

Future Directions

While outcomes from this dissertation are promising, the amount of data used was small. Additionally, results examined were from three very-related intervention programs. To test the robustness of MARS, additional datasets are required. Larger datasets would limit the variability in the data and possibly lessen bias. Additionally, word learning data from other studies that examine different instructional methods would further solidify our findings if results mirrored the results from this dissertation. This would include more collaboration between statisticians and researchers from various disciplines. By creating a symbiotic relationship, statisticians would have access to more "real" data (i.e., not simulated) and researchers would have access to expertise when selecting robust analytic methods.

Once relevant lexical characteristics have been identified, an algorithm will be developed to sequence words for instruction. This algorithm would allow for a systematic approach to vocabulary selection for both teachers and researchers. Users would be able to indicate grade level or age of students, and a list of targets would be provided. Other educationally relevant information could be accessed, such as child-friendly definitions, or examples, and related contexts to use during instruction. This line of research is the first of its kind and has the potential to impact the ways in which we select words for instruction.

In the interim, the MARS word learning regressions can be used to predict the number of students that will learn a given word based on its lexical characteristics. This has the potential to be a powerful tool for researchers and educators to select appropriate words to teach. For example, the words advise, illegal, and space were selected. Their lexical characteristics are listed in Table 7.1. Based on the regressions, we predict that 2% of kindergartners, 21% of first grades, 22% of second graders, and 23% of third graders will learn the word advise. For the word illegal, we predict 11% of kindergartners, 20% of first graders, 34% of second graders, and 40% of third graders will learn the word. We predict 60% of kindergartners, 77% of first graders, 76% of second graders, and 90% of third graders will learn the word space. This information can be used to select vocabulary words appropriate for instruction based on grade level. While it is not the full algorithm, using the MARS regressions provides a basic framework to select vocabulary targets based on grade level. This is the first step in using the relevant lexical characteristics to select vocabulary words for instruction and warrants additional investigation.

Table 7.1. *Example Words with their Lexical Characteristics*

	Age of Acquisition	Level of Concreteness	Word Frequency	Neighborhood Density	Phonotactic Probability
advise	8.89	2.03	12.20	69.83	.09
illegal	9.21	2.37	23.51	38.42	.26
space	5.67	3.54	66.06	32.03	.19

Existing lexical characteristic datasets could be improved upon. For example, word frequency measures do not contain frequency counts for words directly related to children. Using adult frequency measures may not be the most accurate benchmarks when examining children's word learning. Machine learning techniques like natural language processing (NLP) can be used to create a more accurate dataset of word frequencies for children. This can be achieved by focusing on children's literature or subtitle of children's television and movies, mirroring Brysbaert and New's (2009) frequency measures. Once this is created, it can inform the development of other databases relating to lexical characteristics specific to the words children are exposed to. There could be differential effects between adult and child-based norms on children's word learning.

REFERENCES

Agarwal, A., Shah, D., Shen, D., & Song, D. (2021). On robustness of principal component regression. *Journal of the American Statistical Association*, 1-34.

Albanese, D., Filosi, M., Visintainer, R., Riccadonna, S., Jurman, G., & Furlanello, C. (2013). Minerva and minepy: a C engine for the MINE suite and its R, Python and MATLAB wrappers. *Bioinformatics*, *29*(3), 407-408.

Anderson, R. C. & Nagy, W. E. (1991). Word meanings. In R. Barr, M.L. Kamil, P.B Mosenthal, and P. D. Pearson (Eds), *Handbook of reading research*, Vol. 2, 690-724. New York, NY: Longman.

Awad, M., & Khanna, R. (2015). Efficient learning machines: theories, concepts, and applications for engineers and system designers (p. 268). Springer nature.

Bair, E., Hastie, T., Paul, D., & Tibshirani, R. (2006). Prediction by supervised principal components. Journal of the American Statistical Association, 101(473), 119-137.

Baker, S. K., Simmons, D. C., & Kameenui, E. J. (1998). Vocabulary acquisition: Instructional and curricular basics and implications. In D.C. Simmons & E. J. Kameenui (Eds.), *What reading research tells us about children with diverse learning needs: Bases and basics*, 219-238. New York, NY: Routledge.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J., Nelson, D., Simpson, G. & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39(3)*, 445-459.

Bartlett, P., Freund, Y., Lee, W. S., & Schapire, R. E. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics, 26*(5), 1651-1686.

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

Beck, I., McKeown, M. (2007). Increasing young low-income children's oral vocabulary repertoires through rich and focused instruction. *The Elementary School Journal, 107*(3), 251-271.

Beck, I. L., McKeown, M. G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction*. Guilford Press.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity* (Vol. 571). John Wiley & Sons.

Biemiller, A. (2005). Size and sequence in vocabulary development: Implications for choosing words for primary grade vocabulary instruction. In Hiebert, E.H. & Kamil, M.L.(Eds.), *Teaching and learning vocabulary: Bringing Research to Practice* (223-243). Routledge.

Biemiller, A. (2010). *Words worth teaching: Closing the vocabulary gap*. Columbus, OH: McGraw-Hill SRA.

Bishop, C. M. (2006). Pattern recognition. *Machine learning, 128*(9).

Boehmke, B., & Greenwell, B. M. (2019). *Hands-on machine learning with R*. Chaptman and Hall/CRC Press.

Bottou, L., & Lin, C. J. (2007). Support vector machine solvers. *Large Scale Kernel Machines, 3*(1), 301-320.

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, 1287-1294.

Brysbaert, M. & New, B. (2009) Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977-990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46(3)*, 904-911.

Biemiller, A. (2010). *Words worth teaching: Closing the vocabulary gap.* Columbus, OH: McGraw-Hill SRA.

Cao, X., Wang, D., & Wu, L. (2021). Performance of ridge estimator in skew-normal mode regression model. *Communications in Statistics-Simulation and Computation*, 1-14.

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning*, 161-168.

Casella, G., Ghosh, M., Gill, J., & Kyung, M. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis, 5*(2), 369-411.

Cassel, C., Hackl, P., & Westlund, A. H. (1999). Robustness of partial least-squares method for estimating latent variable quality structures. *Journal of Applied Statistics, 26*(4), 435-446.

Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (1999). Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading*, *3*(4), 331-361.

Conover, W. J. (1998). *Practical nonparametric statistics* (Vol. 350). John Wiley & Sons.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Coyne, M. D., McCoach, D. B., & Kapp, S. (2007). Vocabulary intervention for kindergarten students: Comparing extended instruction to embedded instruction and incidental exposure. *Learning Disability Quarterly*, *30*(2), 74-88.

Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, *34*(2), 213-238.

Dale, E., & O'rourke, J. (1976). The Living Word Vocabulary, the Words We Know: A National Vocabulary Inventory.

De Mol, C., De Vito, E., & Rosasco, L. (2009). Elastic-net regularization in learning theory. *Journal of Complexity*, *25*(2), 201-230.

Doornik, J. A., & Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, 70, 927-939.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Münkemüller, T. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27-46.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*, *326*. John Wiley & Sons.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems, 9*, 155-161.

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness*, *2*(1), 1-44.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research, 5*(9).

Francis, W. N., Kucera, H., Kučera, H., & Mackie, A. W. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Houghton Mifflin.

Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics, 35*(2), 109-135.

Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. *International Conference on Machine Learning, 96*, 148-156.

Friedman, J. H. (1991). *Estimating functions of mixed ordinal and categorical variables using adaptive splines*. Stanford Univ CA Lab for Computational Statistics.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, 1-67.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367-378.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of Statistics, 28*(2), 337-407.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software, 33*(1), 1–22. https://www.jstatsoft.org/v33/i01/.

Friedman, J. H., & Roosen, C. B. (1995). An introduction to multivariate adaptive regression splines. *Statistical Methods in Medical Research, 4*(3), 197–217

Friendly, M. (2013). The generalized ridge trace plot: Visualizing bias and precision. *Journal of Computational and Graphical Statistics, 22*(1), 50-68.

German, D. J., & Newman, R. S. (2004). The impact of lexical factors on children's word-finding errors. *Journal of Speech, Language, and Hearing Research, 47*(3), 624-636.

Gierut, J. A., & Dale, R. A. (2007). Comparability of lexical corpora: Word frequency in phonological generalization. *Clinical linguistics & phonetics, 21*(6), 423-433.

Goeman, J., Meijer, R., & Chaturvedi, N. (2018). L1 and L2 penalized regression models. Vignette R Package Penalized. URL http://cran.nedmirror.nl/web/packages/penalized/vignettes/penalized. pdf.

Goldfield, B. A., & Reznick, J. S. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of Child Language*, *17*(1), 171-183.

Goldstein, H., & Kelley, E. S. (2016). Story Friends Teacher Guide. Baltimore: Paul H. Brookes.

Goldstein, H., Ziolkowski, R. A., Bojczyk, K. E., Marty, A., Schneider, N., Harpring, J., & Haring, C. D. (2017). Academic vocabulary learning in first through third grade in low-income schools: Effects of automated supplemental instruction. *Journal of Speech, Language, and Hearing Research, 60*(11), 3237-3258.

Gray, S. (2004). Word learning by preschoolers with Specific Language Impairment: Predictors and poor learners. *Journal of Speech, Language & Hearing Research*, *47*(5).

Gray, S., & Yang, H. C. (2015). Selecting Vocabulary Words to Teach. *SIG 1 Perspectives on Language Learning and Education*, *22*(4), 123-130.

Greenwell, Brando, Boehmke Bradley, Cunningham Jay and GBM Developers (2020). gbm: Generalized Boosted Regression Models. R package version 2.1.8. https://CRAN.R-project.org/package=gbm.

Greenwood, C. R., Carta, J. J., Atwater, J., Goldstein, H., Kaminski, R., & McConnell, S. (2013). Is a response to intervention (RTI) approach to preschool language and early literacy instruction needed? *Topics in Early Childhood Special Education*, *33*(1), 48-64.

Hadley, E. B., Dedrick, R. F., Dickinson, D. K., Kim, E., Hirsh-Pasek, K., & Golinkoff, R. M. (2021). Exploring the relations between child and word characteristics and preschoolers' word-learning. *Journal of Applied Developmental Psychology*, *77*, 101332.

Hadley, E. B., & Mendez, K. Z. (2021). A systematic review of word selection in early childhood vocabulary instruction. *Early Childhood Research Quarterly*, *54*, 44-59.

Hagan, M. T., Demuth, H. B., & Beale, M. (1997). Neural network design. PWS Publishing Co..

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

Hart, B., & Risley, T. R. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator*, *27*(1), 4-9.

Hartmanis, J., & Stearns, R. E. (1965). On the computational complexity of algorithms. *Transactions of the American Mathematical Society*, 117, 285-306.

Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research, 5*(Oct), 1391-1415.

Hebbali, Aravind (2020). olsrr: Tools for Building OLS Regression Models. R package version 0.5.3. https://CRAN.R-project.org/package=olsrr.

Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in statistics-Theory and Methods*, *19*(10), 3595-3617.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics, 12*(1), 69-82.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67.

Hoerl, A. E., & Kennard, R. W. (1976). Ridge regression iterative estimation of the biasing parameter. *Communications in Statistics-Theory and Methods*, *5*(1), 77-88.

Hoover, J. R., Storkel, H. L., & Hogan, T. P. (2010). A cross-sectional comparison of the effects of phonotactic probability and neighborhood density on word learning by preschool children. *Journal of Memory and Language*, *63*(1), 100-116.

Hoque, N., Bhattacharyya, D. K., & Kalita, J. K. (2014). MIFS-ND: A mutual information-based feature selection method. *Expert Systems with Applications, 41*(14), 6371-6385.

Huber, P. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics 53*, 73-101.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112), 18. New York: Springer.

Johnson, R. A., & Wichern, D. W. (2014). *Applied multivariate statistical analysis (Vol. 6).* London, UK: Pearson.

Ju, X., Chen, V. C., Rosenberger, J. M., & Liu, F. (2021). Fast knot optimization for multivariate adaptive regression splines using hill climbing methods. *Expert Systems with Applications, 171*, 114565.

Justice, L. M., Meier, J., & Walpole, S. (2005). Learning new words from storybooks: An efficacy study with at-risk kindergartners. *Language, Speech, and Hearing Services in Schools, 36*(1), 17-32.

Kearns, M. J. (1990). *The computational complexity of machine learning*. MIT press.

Kelley, E. S., Barker, R. M., Peters-Sanders, L., Madsen, K., Seven, Y., Soto, X., Olsen, W., Hull, K., & Goldstein, H. (2020). Feasible implementation strategies for improving vocabulary knowledge of high-risk preschoolers: Results from a cluster-randomized trial. *Journal of Speech, Language, and Hearing Research*, *63*(12), 4000-4017.

Kilic Depren, S. (2018). Prediction of Students' Science Achievement: An Application of Multivariate Adaptive Regression Splines and Regression Trees. *Journal of Baltic Science Education, 17*(5), 887.

Knuth, D. E. (1976). Big omicron and big omega and big theta. *ACM Sigact News, 8*(2), 18-24.

Kolyshkina, I., Brownlow, M., & Taylor, J. (2013, December). Improving every child's chance in life. In *2013 IEEE 13th International Conference on Data Mining Workshops* (pp. 180-184). IEEE.

Korkmaz S, Goksuluk D, Zararsiz G. MVN: (2014). An R package for assessing multivariate normality. *The R Journal*,*6*(2):151-162.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44(4)*, 978-990.

Kvålseth, T. O. (1985). Cautionary note about R 2. *The American Statistician*, *39*(4), 279-285.

Liaw, A. and Wiener M. (2002). Classification and Regression by random forest. *R News 2*(3), 18--22.

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *Annals of Statistics*, *42*(2), 413.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19(1)*, 1-36.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519-530.

Martís, R., Alonso, J., Catalán, C., Fuentes, R., & Suárez, A. A. (2015, August). Prediction of the Student Success Rate by means of Quality Teaching Survey Variables Applying a Multivariate Adaptive Regression Splines (Mars) Models. In *Toulon-Verona Conference "Excellence in Services"*.

Marzano, R. J., & Pickering, D. J. (2005). *Building academic vocabulary: Teacher's manual*. Association for Supervision and Curriculum Development.

Marzano, R. J., & Simms, J. A. (2013). *Vocabulary for the Common Core*. Bloomington, IN: Marzano Research Laboratory.

Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association, 60*(309), 234-256.

McDonough, C., Song, L., Hirsh-Pasek, K., Golinkoff, R. M., & Lannon, R. (2011). An image is worth a thousand words: Why nouns tend to dominate verbs in early word learning. *Developmental Science*, *14*(2), 181-189.

McKeown, M. G., & Beck, I. L. (2014). Effects of vocabulary instruction on measures of language processing: Comparing two approaches. *Early childhood Research quarterly*, *29*(4), 520-530.

Meyer, D., Dimitriadou, E. Hornik, K., Weingessel, A., and Leisch, F (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. *R package version 1*(3). https://CRAN.R-project.org/package=e1071.

Meyer, D., & Wien, F. T. (2015). Support vector machines. *The Interface to libsvm in package e1071, 28.*

Milborrow, M. S. (2011). Derived from mda:mars by T. Hastie and R. Tibshirani. *earth: Multivariate Adaptive Regression Splines*. R package.

Morrison, C. M., & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology*, *91*(2), 167-180.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Muniz, G., & Kibria, B. G. (2009). On some ridge regression estimators: An empirical comparison. *Communications in Statistics—Simulation and Computation, 38*(3), 621-630.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, *7*, 21.

National Reading Panel. (2000). Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Washington, DC: National Institute of Child Health and Human Development.

Newman, R. S., & German, D. J. (2002). Effects of lexical factors on lexical access among typical language-learning children and children with word-finding difficulties. *Language and Speech*, *45*(3), 285-317.

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, and Evaluation*, *8*(1), 2.

Ponari, M., Norbury, C. F., & Vigliocco, G. (2018). Acquisition of abstract concepts is influenced by emotional valence. *Developmental Science*, *21*(2), e12549.

R Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org/.

Ramachandran, K. M., & Tsokos, C. P. (2020). *Mathematical statistics with applications in R*. Academic Press.

Rencher, A. C., & Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., ... & Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, *334*(6062), 1518-1524.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., & Ripley, M. B. (2013). Package 'mass'. *Cran r*, *538*, 113-120.

Rosipal, R. (2011). Nonlinear partial least squares an overview. In: *Chemoinformatics and advanced machine learning perspectives: complex computational methods and collaborative techniques, 169-189.*

Rosipal, R., & Krämer, N. (2005, February). Overview and recent advances in partial least squares. In International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection" (pp. 34-51). Springer, Berlin, Heidelberg.

Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing*, *2*(3), 117-119.

Sakamoto, W. (2007). MARS: selecting basis functions and knots with an empirical Bayes method. *Computational Statistics*, *22*(4), 583-597.

Scarborough, H. S. (2001). Connecting early language and literacy to later reading (dis)abilities: Evidence, theory, and practice. In S. Neuman, & D. Dickinson (Eds.), *Handbook for research in early literacy* (pp.97-110). New York: Guilford Press.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197-227.

Schapire, R. E., & Freund, Y. (2013). Boosting: Foundations and algorithms. Kybernetes.

Sénéchal, M., Ouellette, G., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. *Handbook of early literacy research*, *2*, 173-182.

Shutes, K., & Adcock, C. (2013). Regularized Extended Skew-Normal Regression.

Šíma, J., & Orponen, P. (2003). General-purpose computation with neural networks: A survey of complexity theoretic results. *Neural Computation*, *15*(12), 2727-2778.

Smola, A. J. (1996). Regression estimation with support vector learning machines (Doctoral dissertation, Master's thesis, Technische Universität München).

Smola, A. J., & Schölkopf, B. (1998). Learning with kernels (Vol. 4). GMD-Forschungszentrum Informationstechnik.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199-222.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). Preventing reading difficulties in young children committee on the prevention of reading difficulties in young children. *Washington, DC: National Research Council*.

Snyder, P., Thompson, B., Mclean, M. E., & Smith, B. J. (2002). Examination of quantitative methods used in early intervention research: Linkages with recommended practices. *Journal of Early Intervention*, *25*(2), 137-150.

Song, L., Vempala, S., Wilmes, J., & Xie, B. (2017). On the complexity of learning neural networks. arXiv preprint arXiv:1707.04615.

Stahl, S. A., & Nagy, W. E. (2007). *Teaching word meanings*. Routledge.

Stoel-Gammon, C. (2011). Relationships between lexical and phonological development in young children. *Journal of child language*, *38*(1), 1-34.

Stone, M., & Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, *52*(2), 237-258.

Storkel, H. (2001). Learning new words: phonotactic probability in language development. *Journal of Speech, Language, and Hearing Research, 44*(6), 1321-1377. (doi:10.1044/1092-4388(2001/103).

Storkel, H. L. (2004). Methods for minimizing the confounding effects of word length in the analysis of phonotactic probability and neighborhood density. *Journal of Speech, Language, and Hearing Research, 47(6),* 1454-1468.

Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical, and semantic variables on word learning by infants. *Journal of Child Language, 36*(2), 291.

Storkel, H. L., Armbrüster, J., & Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research, 49(6),* 1175-1192. Chicago.

Storkel, H. L., & Rogers, M. A. (2000). The effect of probabilistic phonotactics on lexical acquisition. *Clinical Linguistics & Phonetics, 14*(6), 407-425.

Storkel, H. L., Voelmle, K., Fierro, V., Flake, K., Fleming, K. K., & Romine, R. S. (2017). Interactive book reading to accelerate word learning by kindergarten children with specific language impairment: Identifying an adequate intensity and variation in treatment response. *Language, Speech, and Hearing Services in Schools, 48*(1), 16-30.

Székely, G. J., & Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis, 93*(1), 58-80.

Taffe, S. W., Blachowicz, C. L. Z., & Fisher, P. J. (2009). Vocabulary instruction for diverse students. In L.M. Morrow, R. Rueda, & D. Lapp (Eds.), *Handbook of research on literacy & diversity*, (pp. 320-336). New York: Guilford Press.

Tamhane, A., & Dunlop, D. (2000). *Statistics and data analysis: from elementary to intermediate*. Upper Saddle River, NJ: Prentice Hall, Inc.

Therneau, Terry and Atkinson, Beth (2019). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-15. https://CRAN.R-project.org/package=rpart.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267-288.

Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungs problem. *Proceedings of the London Mathematical Society, 2*(1), 230-265.

Vapnik, V. (2013). *The nature of statistical learning theory.* Springer science & business media.

Vaden, K. I., Halpin, H. R., & Hickok, G. S. (2009). Irvine Phonotactic Online Dictionary, Version 2.0 [Data file]. Available at www.iphod.com.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S.* Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Vitevitch, M. S., Armbrüster, J., & Chu, S. (2004). Sublexical and lexical representations in speech production: effects of phonotactic probability and onset density. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(2), 514.

Vitevitch, M.S. & Luce, P.A. (2004) A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers,* 36, 481-487.

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language, 68*(1-2), 306-311.

Wasik, B. A., Hindman, A. H., & Snell, E. K. (2016). Book reading and vocabulary development: A systematic review. *Early Childhood Research Quarterly, 37*, 39–57.

Wehrens, R., & Mevik, B. H. (2007). The pls package: principal component and partial least squares regression in R.

Were, K., Bui, D. T., Dick, Ø. B., & Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators, 52,* 394-403.

Willett, J. B., & Singer, J. D. (1988). Another cautionary note about R 2: Its use in weighted least-squares regression analysis. *The American Statistician, 42*(3), 236-238.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research, 30*(1), 79-82.

Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability, 12*(S1), 117-142.

Wright, T. S. (2012). What classroom observations reveal about oral vocabulary instruction in kindergarten. *Reading Research Quarterly, 47*(4), 353-355.

Yu, C. H., Digangi, S., Jannasch-Pennell, A. K., & Kaprolet, C. (2008). Profiling students who take online courses using data mining methods. *Online Journal of Distance Learning Administration, 11*(2), 1-14.

Yu, C. H., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, *8*(2), 307-325.

Zaffalon, M., & Hutter, M. (2002). Robust feature selection by mutual information distributions. arXiv preprint cs/0206006.

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. R News 2: 7–10. Available at (accessed August 2011). http://CRAN. R-project. org/doc/Rnews/(http://CRAN. R-project. org/doc/Rnews/).

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.

Zou, K. H., Tuncali, K., & Silverman, S. G. (2003). Correlation and simple linear regression. *Radiology*, *227*(3), 617-628.

APPENDIX I:

STEP BY STEP GUIDE FOR APPLYING MARS

Introduction and Descriptive Statistics

Before applying multivariate adaptive regression splines, the data needs to be loaded and tested for multicollinearity, homoscedasticity, univariate normality, and multivariate normality. The packages that will be used are "e1071" to test skewness, "olsrr" for VIF, "minerva" for MIC, "MVN" for multivariate normal tests, "stats" for correlations, "corrplot" to plot correlations, "earth" to build the MARS model, and "Metrics" to check model fit.

The first step is to load the data. Begin by setting the working directory with the setwd() command. In the example given, the default file location will be D:/Research. If the data is in CSV form, we use read.csv to load the data from the filepath D:/Research/ILIAD/1stGrade/ILIAD1stGradeClean.csv for this example and name the dataset data1.

```
setwd("D:/Research")

#load data - Example uses ILIAD first-grade data
data1 <- read.csv(file="D:/Research/ILIAD/1stGrade/ILIAD1stGradeClean.csv", h
eader=TRUE)
data1 <- data1[complete.cases(data1), ]
```

Next we will check the descriptive statistics for each independent variable. For the ILIAD data, these include word frequency (SUBtLwf), age of acquisition (AoA), level of concreteness (Conc_Mean), neighborhood density (N_Den), and phonotactic probability (Phon_Prob). Mean,

standard deviation, median, minimum value, maximum value, and skewness will be measured for each parameter.

Word Frequency

```
#load library
#If the package needs to be installed use the install.packages command
#install.packages("e1071", dep=TRUE)
library(e1071) #Skewness

#descriptives for word frequency
mean(data1$SUBTLwf) #mean

## [1] 19.28967

sd(data1$SUBTLwf) #standard deviation

## [1] 57.93981

median(data1$SUBTLwf) #median

## [1] 6.9

min(data1$SUBTLwf) #minimum value

## [1] 0.27

max(data1$SUBTLwf) #maximum value

## [1] 509.37

skewness(data1$SUBTLwf) #skew

## [1] 6.672805
```

Age of Acquisition

```
#age of acquisition
mean(data1$AoA)

## [1] 8.799706

sd(data1$AoA)

## [1] 2.172357

median(data1$AoA)

## [1] 9.06

min(data1$AoA)

## [1] 3.25
```

```
max(data1$AoA)

## [1] 13.61

skewness(data1$AoA)

## [1] -0.4591858
```

Level of Concreteness

```
#level of concreteness
mean(data1$Conc_Mean)

## [1] 2.955245

sd(data1$Conc_Mean)

## [1] 0.9725028

median(data1$Conc_Mean)

## [1] 2.76

min(data1$Conc_Mean)

## [1] 1.5

max(data1$Conc_Mean)

## [1] 5

skewness(data1$Conc_Mean)

## [1] 0.6379685
```

Neighborhood Density

```
#neighborhood density
mean(data1$N_Den)

## [1] 1845.324

sd(data1$N_Den)

## [1] 8474.12

median(data1$N_Den)

## [1] 6.49

min(data1$N_Den)

## [1] 0

max(data1$N_Den)
```

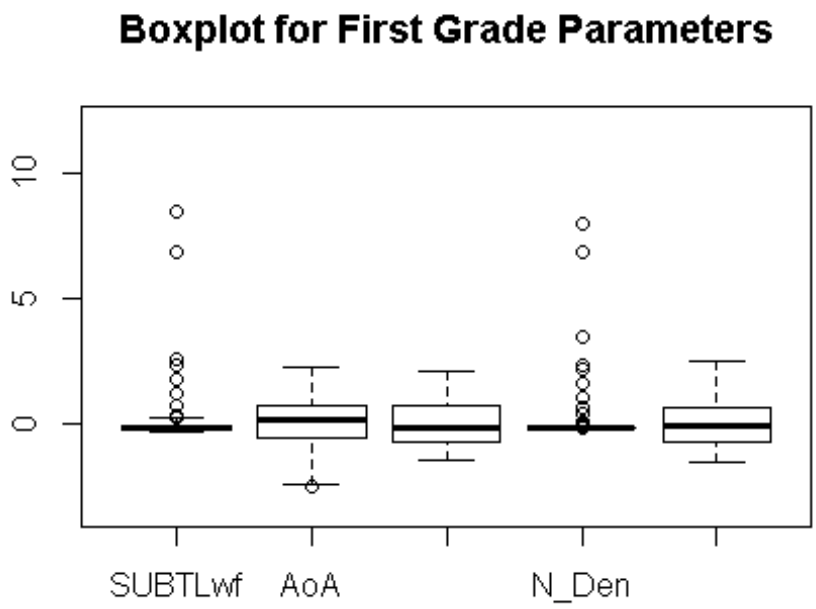
```
## [1] 69210.62

skewness(data1$N_Den)

## [1] 6.232289
```

Phonotactic Probability

```
#phonotactic probability
mean(data1$Phon_Prob)

## [1] 0.2211839

sd(data1$Phon_Prob)

## [1] 0.1217351

median(data1$Phon_Prob)

## [1] 0.2058

min(data1$Phon_Prob)

## [1] 0.0332

max(data1$Phon_Prob)

## [1] 0.5176

skewness(data1$Phon_Prob)

## [1] 0.5657482
```

A boxplot for the parameters can be created using the boxplot command. The scale() will scale the data so that it can be compared more easily; otherwise neighborhood density will overshadow the other variables because it has such a wide range of large values. Within scale() the data is selected, columns 3 through 7 in this case. For the ILIAD data, column 1 is the decontextualized learning as a decimal, and column 2 is the expressive task learning as a decimal. ylim controls the upper and lower limits for the y-axis.

```
boxplot(scale(data1[,3:7]), main="Boxplot for First Grade Parameters", ylim=c
(-3.5,12))
```



Exploring the Data

For the ILIAD data, the descriptive statistics have shown that neighborhood density and word frequency are highly skewed. This can lead to unreliable models. Next we will perform an exploratory data analysis. We will begin by looking at variance inflation factor and maximal information coefficient. Both of these are measurements of multicollinearity. There is no definitive threshold for multicollinearity using VIF, but generally anything larger than 4 is considered to have moderately high multicollinearity and above 10 is very high multicollinearity. MIC is an information theory technique that can find both linear and non-linear relationships within the data.

```
#VIF and MIC
#instell.packages("olsrr", dep=TRUE)
library(olsrr) #VIF
#instell.packages("minerva", dep=TRUE)
library(minerva) #MIC
lmMod1 <- lm(Decon ~ SUBTLwf+AoA+Conc_Mean+N_Den+Phon_Prob, data=data1) #line
```

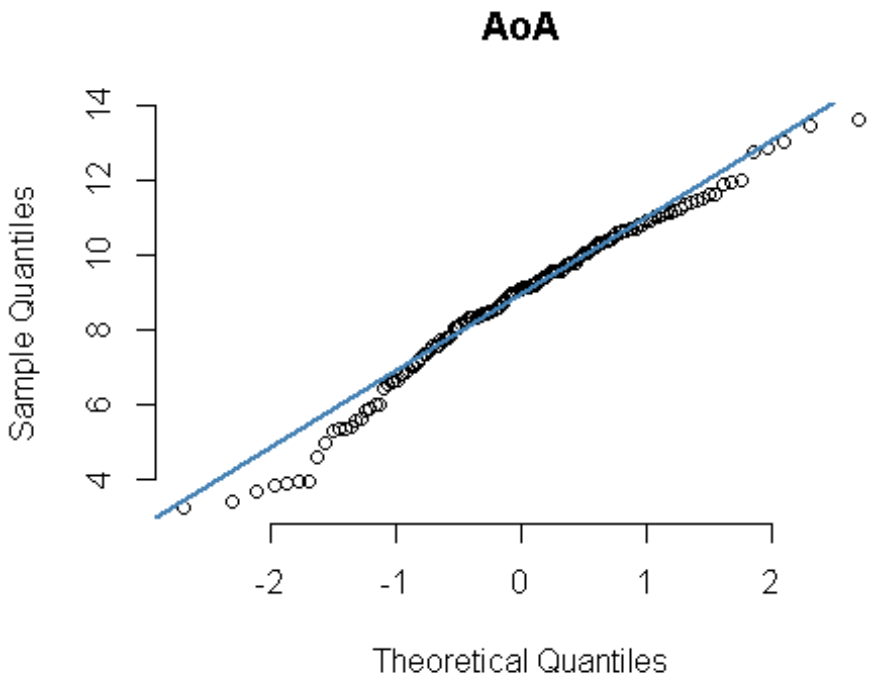
```
ar model for VIF calculations
ols_vif_tol(lmMod1) #Calculate VIF

##   Variables Tolerance      VIF
## 1   SUBTLwf 0.8625632 1.159335
## 2       AoA 0.5452588 1.833991
## 3 Conc_Mean 0.6119911 1.634011
## 4     N_Den 0.9251188 1.080942
## 5 Phon_Prob 0.8527566 1.172668

mine(data1[,3:7], measure="mic") #MIC

## $MIC
##             SUBTLwf       AoA Conc_Mean     N_Den Phon_Prob
## SUBTLwf   0.9999647 0.3514268 0.2540276 0.3328488 0.2210946
## AoA       0.3514268 0.9999647 0.4018248 0.3530997 0.2438121
## Conc_Mean 0.2540276 0.4018248 0.9999647 0.2681356 0.2541717
## N_Den     0.3328488 0.3530997 0.2681356 0.9999647 0.2776106
## Phon_Prob 0.2210946 0.2438121 0.2541717 0.2776106 0.9999647
##
## $MAS
##              SUBTLwf        AoA  Conc_Mean      N_Den  Phon_Prob
## SUBTLwf   0.00000000 0.06660012 0.02707641 0.05348317 0.03832547
## AoA       0.06660012 0.00000000 0.05967042 0.08537223 0.02396344
## Conc_Mean 0.02707641 0.05967042 0.00000000 0.04618807 0.02385355
## N_Den     0.05348317 0.08537223 0.04618807 0.00000000 0.06236391
## Phon_Prob 0.03832547 0.02396344 0.02385355 0.06236391 0.00000000
##
## $MEV
##             SUBTLwf       AoA Conc_Mean     N_Den Phon_Prob
## SUBTLwf   0.9999647 0.3514268 0.2540276 0.3328488 0.2210946
## AoA       0.3514268 0.9999647 0.4018248 0.3530997 0.2438121
## Conc_Mean 0.2540276 0.4018248 0.9999647 0.2681356 0.2541717
## N_Den     0.3328488 0.3530997 0.2681356 0.9999647 0.2776106
## Phon_Prob 0.2210946 0.2438121 0.2541717 0.2776106 0.9999647
##
## $MCN
##           SUBTLwf      AoA Conc_Mean    N_Den Phon_Prob
## SUBTLwf         2 2.000000         2 2.000000         2
## AoA             2 2.000000         2 2.584963         2
## Conc_Mean       2 2.000000         2 2.000000         2
## N_Den           2 2.584963         2 2.000000         2
## Phon_Prob       2 2.000000         2 2.000000         2
##
## $MICR2
##                SUBTLwf           AoA     Conc_Mean         N_Den     Phon
_Prob
## SUBTLwf   -3.527573e-05  2.203353e-01  1.813489e-01  3.325425e-01  2.05180
7e-01
## AoA        2.203353e-01 -3.527573e-05  1.880936e-02  3.494168e-01  1.58174
```

```
3e-01
## Conc_Mean  1.813489e-01  1.880936e-02 -3.527573e-05  2.616560e-01  2.24882
0e-01
## N_Den      3.325425e-01  3.494168e-01  2.616560e-01 -3.527573e-05  2.09067
9e-01
## Phon_Prob  2.051807e-01  1.581743e-01  2.248820e-01  2.090679e-01 -3.52757
3e-05
##
## $GMIC
##              SUBTLwf       AoA Conc_Mean     N_Den Phon_Prob
## SUBTLwf   0.9999647 0.2611213 0.1722910 0.2489660 0.1320849
## AoA       0.2611213 0.9999647 0.3204352 0.2569411 0.1744159
## Conc_Mean 0.1722910 0.3204352 0.9999647 0.1991796 0.1739194
## N_Den     0.2489660 0.2569411 0.1991796 0.9999647 0.2059763
## Phon_Prob 0.1320849 0.1744159 0.1739194 0.2059763 0.9999647
##
## $TIC
##              SUBTLwf       AoA Conc_Mean     N_Den Phon_Prob
## SUBTLwf   22.999105  5.084633  3.523779  4.868565  2.737394
## AoA        5.084633 22.998965  6.595474  4.662306  3.761417
## Conc_Mean  3.523779  6.595474 22.999105  3.703152  3.678444
## N_Den      4.868565  4.662306  3.703152 22.999105  3.902378
## Phon_Prob  2.737394  3.761417  3.678444  3.902378 22.999105
```

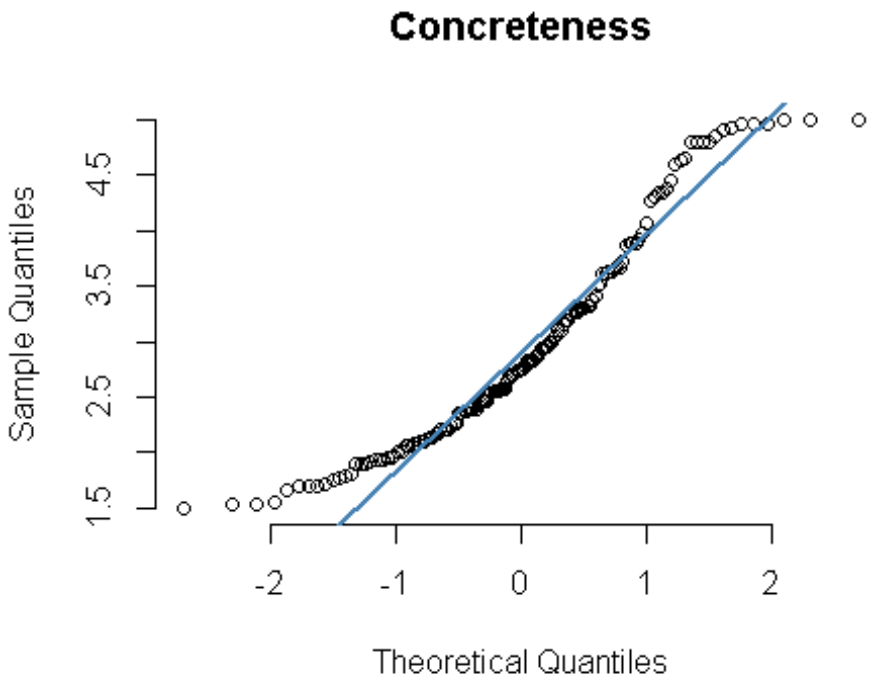
Univariate Normality

Next, we will check univariate normality (whether each variable follows a normal distribution) using Q-Q plots and Shapiro Wilk W Test for Univariate Normal.

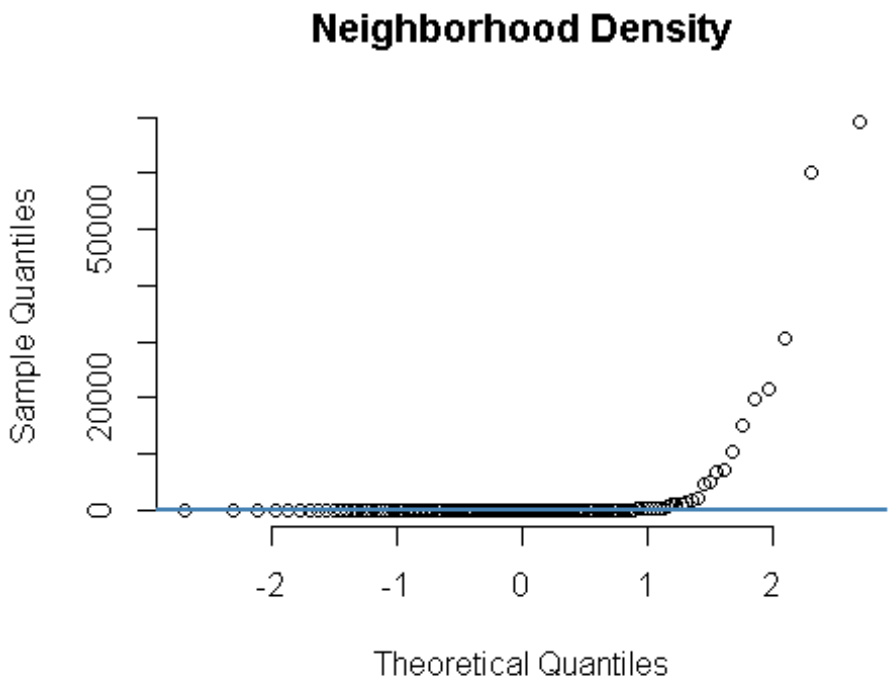
```
#qqplots for each variable
qqnorm(data1$AoA, pch = 1, frame = FALSE, main="AoA")
qqline(data1$AoA, col = "steelblue", lwd = 2)
```



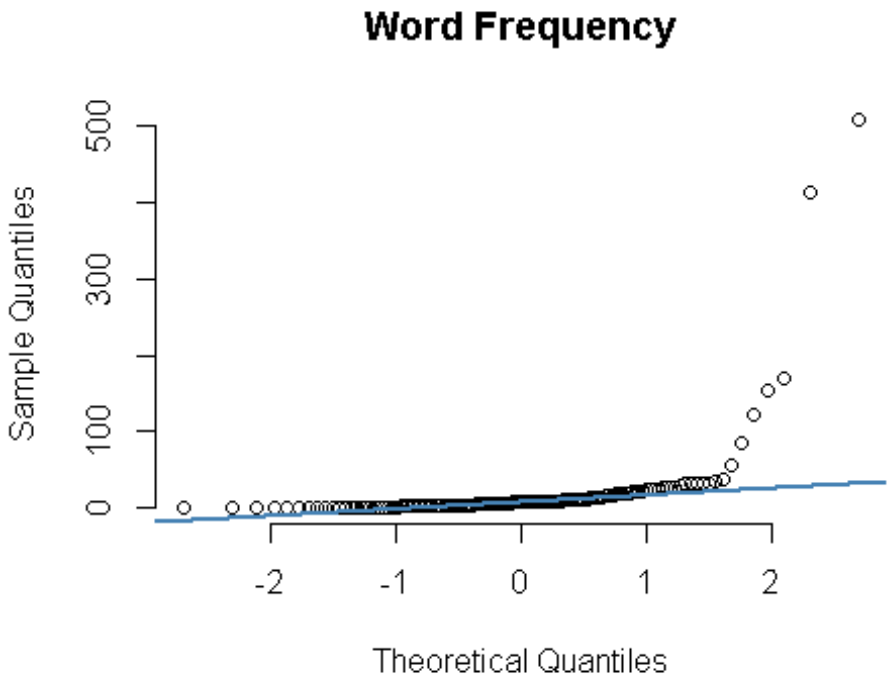

```
qqnorm(data1$Conc_Mean, pch = 1, frame = FALSE, main="Concreteness")
qqline(data1$Conc_Mean, col = "steelblue", lwd = 2)
```



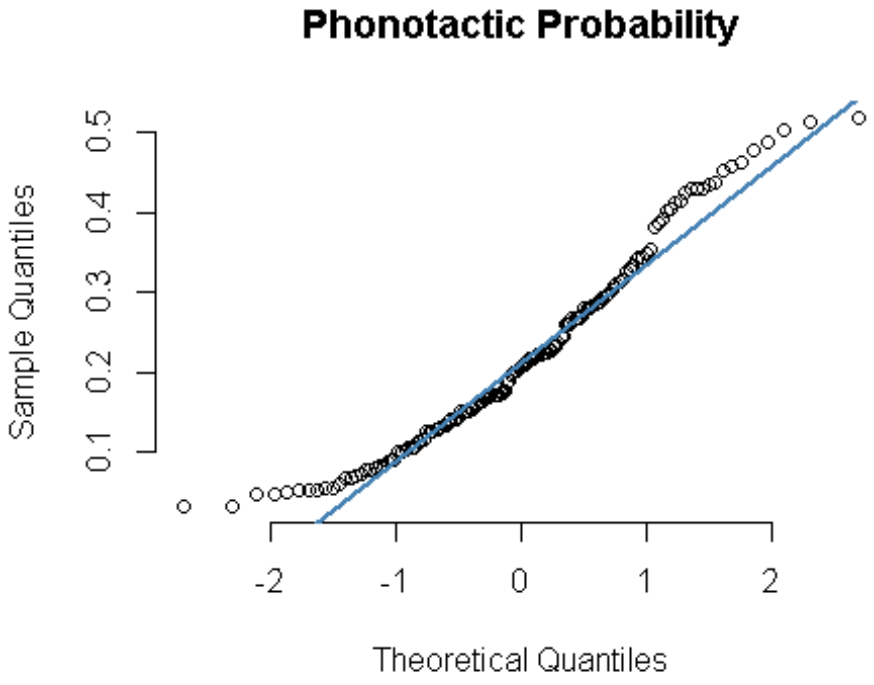
```
qqnorm(data1$N_Den, pch = 1, frame = FALSE, main="Neighborhood Density")
qqline(data1$N_Den, col = "steelblue", lwd = 2)
```




```
qqnorm(data1$SUBTLwf, pch = 1, frame = FALSE, main="Word Frequency")
qqline(data1$SUBTLwf, col = "steelblue", lwd = 2)
```




```
qqnorm(data1$Phon_Prob, pch = 1, frame = FALSE, main="Phonotactic Probability
")
qqline(data1$Phon_Prob, col = "steelblue", lwd = 2)
```



```
#Shapiro Wilk W Test for Univariate Normal for each variable
shapiro.test(data1$AoA)

##
##  Shapiro-Wilk normality test
##
## data:  data1$AoA
## W = 0.97709, p-value = 0.01682

shapiro.test(data1$Conc_Mean)

##
##  Shapiro-Wilk normality test
##
## data:  data1$Conc_Mean
## W = 0.92973, p-value = 1.587e-06

shapiro.test(data1$N_Den)

##
##  Shapiro-Wilk normality test
##
## data:  data1$N_Den
## W = 0.22723, p-value < 2.2e-16

shapiro.test(data1$Phon_Prob)

##
##  Shapiro-Wilk normality test
##
## data:  data1$Phon_Prob
## W = 0.95288, p-value = 8.65e-05

shapiro.test(data1$SUBTLwf)

##
##  Shapiro-Wilk normality test
##
## data:  data1$SUBTLwf
## W = 0.28031, p-value < 2.2e-16
```

Multivariate Normality

The normality of the entire dataset, or multivariate normality, can be tested in multiple ways. Using the "MVN" package, we will test this using Mardia Skewness and Kurtosis, Henze-Zirkler test, Royston's H test, Doornik-Hansen's test, and the energy E-statistic.

```
#library.packages("MVN", dep=TRUE)
library(MVN) #Multivariate Normal Tests
```

```r
#Multivariate Normal Tests
mvn(data1[,3:7], mvnTest = "mardia")
```

```
## $multivariateNormality
##              Test        Statistic p value Result
## 1 Mardia Skewness 2130.56804492719       0     NO
## 2 Mardia Kurtosis 61.9953944148249       0     NO
## 3             MVN             <NA>    <NA>     NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk  SUBTLwf     0.2803  <0.001      NO
## 2 Shapiro-Wilk    AoA       0.9771  0.0168      NO
## 3 Shapiro-Wilk Conc_Mean    0.9297  <0.001      NO
## 4 Shapiro-Wilk   N_Den      0.2272  <0.001      NO
## 5 Shapiro-Wilk Phon_Prob    0.9529   1e-04      NO
##
## $Descriptives
##             n         Mean      Std.Dev Median    Min         Max    25th
75th
## SUBTLwf   143   19.2896709   57.9398068 6.9000 0.2700   509.3700 2.38000 1
4.50912
## AoA       143    8.7997064    2.1723573 9.0600 3.2500    13.6100 7.57000 1
0.33500
## Conc_Mean 143    2.9552448    0.9725028 2.7600 1.5000     5.0000 2.17500
3.61500
## N_Den     143 1845.3240559 8474.1200704 6.4900 0.0000 69210.6200 0.70000 5
8.48000
## Phon_Prob 143    0.2211839    0.1217351 0.2058 0.0332     0.5176 0.12875
0.29375
##                Skew    Kurtosis
## SUBTLwf    6.6728053 48.30658409
## AoA       -0.4591858 -0.03939961
## Conc_Mean  0.6379685 -0.62283930
## N_Den      6.2322893 41.77014477
## Phon_Prob  0.5657482 -0.53091382
```

```r
mvn(data1[,3:7], mvnTest = "hz")
```

```
## $multivariateNormality
##            Test       HZ p value MVN
## 1 Henze-Zirkler 7.273224       0  NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk  SUBTLwf     0.2803  <0.001      NO
## 2 Shapiro-Wilk    AoA       0.9771  0.0168      NO
## 3 Shapiro-Wilk Conc_Mean    0.9297  <0.001      NO
## 4 Shapiro-Wilk   N_Den      0.2272  <0.001      NO
## 5 Shapiro-Wilk Phon_Prob    0.9529   1e-04      NO
```

```
##
## $Descriptives
##             n         Mean      Std.Dev Median    Min        Max    25th
75th
## SUBTLwf   143   19.2896709   57.9398068 6.9000 0.2700   509.3700 2.38000 1
4.50912
## AoA       143    8.7997064    2.1723573 9.0600 3.2500    13.6100 7.57000 1
0.33500
## Conc_Mean 143    2.9552448    0.9725028 2.7600 1.5000     5.0000 2.17500
3.61500
## N_Den     143 1845.3240559 8474.1200704 6.4900 0.0000 69210.6200 0.70000 5
8.48000
## Phon_Prob 143    0.2211839    0.1217351 0.2058 0.0332     0.5176 0.12875
0.29375
##                 Skew    Kurtosis
## SUBTLwf    6.6728053 48.30658409
## AoA       -0.4591858 -0.03939961
## Conc_Mean  0.6379685 -0.62283930
## N_Den      6.2322893 41.77014477
## Phon_Prob  0.5657482 -0.53091382

mvn(data1[,3:7], mvnTest = "royston")

## $multivariateNormality
##      Test        H      p value MVN
## 1 Royston 231.6819 2.441715e-48  NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk  SUBTLwf     0.2803  <0.001      NO
## 2 Shapiro-Wilk    AoA       0.9771  0.0168      NO
## 3 Shapiro-Wilk Conc_Mean    0.9297  <0.001      NO
## 4 Shapiro-Wilk   N_Den      0.2272  <0.001      NO
## 5 Shapiro-Wilk Phon_Prob    0.9529   1e-04      NO
##
## $Descriptives
##             n         Mean      Std.Dev Median    Min        Max    25th
75th
## SUBTLwf   143   19.2896709   57.9398068 6.9000 0.2700   509.3700 2.38000 1
4.50912
## AoA       143    8.7997064    2.1723573 9.0600 3.2500    13.6100 7.57000 1
0.33500
## Conc_Mean 143    2.9552448    0.9725028 2.7600 1.5000     5.0000 2.17500
3.61500
## N_Den     143 1845.3240559 8474.1200704 6.4900 0.0000 69210.6200 0.70000 5
8.48000
## Phon_Prob 143    0.2211839    0.1217351 0.2058 0.0332     0.5176 0.12875
0.29375
##                 Skew    Kurtosis
## SUBTLwf    6.6728053 48.30658409
```

```
## AoA       -0.4591858 -0.03939961
## Conc_Mean  0.6379685 -0.62283930
## N_Den      6.2322893 41.77014477
## Phon_Prob  0.5657482 -0.53091382
```

```
mvn(data1[,3:7], mvnTest = "dh")
```

```
## $multivariateNormality
##             Test        E df p value MVN
## 1 Doornik-Hansen 1751.036 10       0  NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk  SUBTLwf     0.2803  <0.001      NO
## 2 Shapiro-Wilk    AoA       0.9771  0.0168      NO
## 3 Shapiro-Wilk Conc_Mean    0.9297  <0.001      NO
## 4 Shapiro-Wilk   N_Den      0.2272  <0.001      NO
## 5 Shapiro-Wilk Phon_Prob    0.9529   1e-04      NO
##
## $Descriptives
##             n         Mean      Std.Dev Median    Min         Max    25th
75th
## SUBTLwf   143   19.2896709   57.9398068 6.9000 0.2700   509.3700 2.38000 1
4.50912
## AoA       143    8.7997064    2.1723573 9.0600 3.2500    13.6100 7.57000 1
0.33500
## Conc_Mean 143    2.9552448    0.9725028 2.7600 1.5000     5.0000 2.17500
3.61500
## N_Den     143 1845.3240559 8474.1200704 6.4900 0.0000 69210.6200 0.70000 5
8.48000
## Phon_Prob 143    0.2211839    0.1217351 0.2058 0.0332     0.5176 0.12875
0.29375
##                Skew    Kurtosis
## SUBTLwf    6.6728053 48.30658409
## AoA       -0.4591858 -0.03939961
## Conc_Mean  0.6379685 -0.62283930
## N_Den      6.2322893 41.77014477
## Phon_Prob  0.5657482 -0.53091382
```

```
mvn(data1[,3:7], mvnTest = "energy")
```

```
## $multivariateNormality
##          Test Statistic p value MVN
## 1 E-statistic  11.86507       0  NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk  SUBTLwf     0.2803  <0.001      NO
## 2 Shapiro-Wilk    AoA       0.9771  0.0168      NO
## 3 Shapiro-Wilk Conc_Mean    0.9297  <0.001      NO
## 4 Shapiro-Wilk   N_Den      0.2272  <0.001      NO
```

```
## 5 Shapiro-Wilk Phon_Prob    0.9529   1e-04      NO
##
## $Descriptives
##              n         Mean      Std.Dev Median    Min         Max    25th
75th
## SUBTLwf   143   19.2896709   57.9398068 6.9000 0.2700   509.3700 2.38000 1
4.50912
## AoA       143    8.7997064    2.1723573 9.0600 3.2500    13.6100 7.57000 1
0.33500
## Conc_Mean 143    2.9552448    0.9725028 2.7600 1.5000     5.0000 2.17500
3.61500
## N_Den     143 1845.3240559 8474.1200704 6.4900 0.0000 69210.6200 0.70000 5
8.48000
## Phon_Prob 143    0.2211839    0.1217351 0.2058 0.0332     0.5176 0.12875
0.29375
##                 Skew    Kurtosis
## SUBTLwf    6.6728053 48.30658409
## AoA       -0.4591858 -0.03939961
## Conc_Mean  0.6379685 -0.62283930
## N_Den      6.2322893 41.77014477
## Phon_Prob  0.5657482 -0.53091382
```

Correlation Between Variables

Correlation between variables will be checked using the Pearson correlation coefficient, Kendall rank correlation, and Spearman's rank correlation.

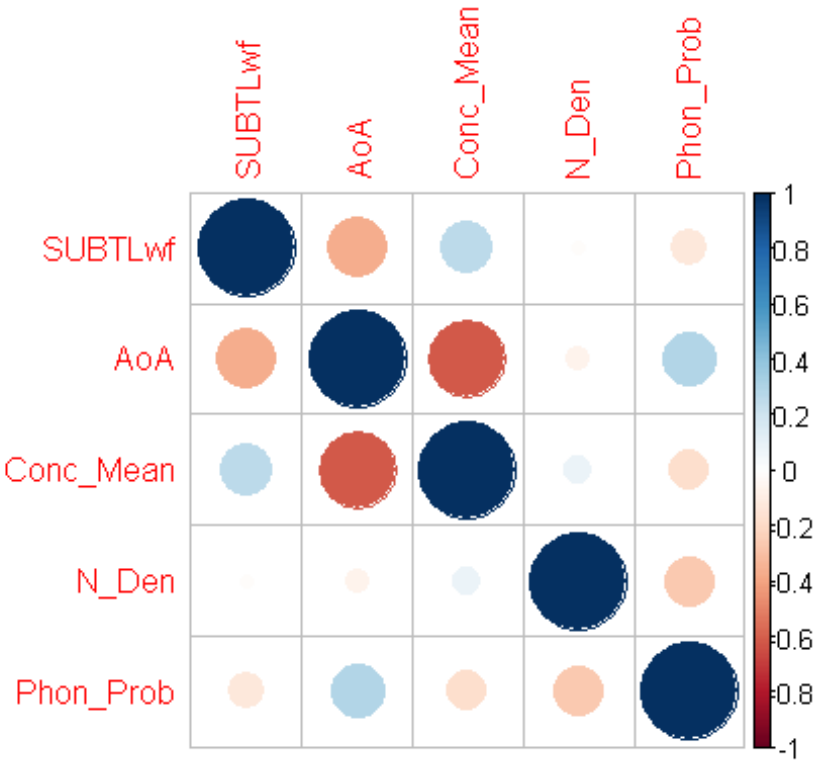
```
#instell.packages("stats", dep=TRUE)
library(stats) #Correlations
corrPearson <- cor(data1[,3:7], method = "pearson")
round(corrPearson,2)

##           SUBTLwf   AoA Conc_Mean N_Den Phon_Prob
## SUBTLwf      1.00 -0.36      0.27 -0.02     -0.13
## AoA         -0.36  1.00     -0.62 -0.06      0.29
## Conc_Mean    0.27 -0.62      1.00  0.08     -0.17
## N_Den       -0.02 -0.06      0.08  1.00     -0.26
## Phon_Prob   -0.13  0.29     -0.17 -0.26      1.00

corrKendall <- cor(data1[,3:7], method = "kendall")
round(corrKendall,2)

##           SUBTLwf   AoA Conc_Mean N_Den Phon_Prob
## SUBTLwf      1.00 -0.37      0.21  0.28     -0.02
## AoA         -0.37  1.00     -0.40 -0.29      0.21
## Conc_Mean    0.21 -0.40      1.00  0.26     -0.11
## N_Den        0.28 -0.29      0.26  1.00     -0.26
## Phon_Prob   -0.02  0.21     -0.11 -0.26      1.00
```

```
corrSpearman <- cor(data1[,3:7], method = "spearman")
round(corrSpearman,2)
```

```
##            SUBTLwf   AoA Conc_Mean N_Den Phon_Prob
## SUBTLwf       1.00 -0.52      0.32  0.41     -0.03
## AoA          -0.52  1.00     -0.56 -0.41      0.31
## Conc_Mean     0.32 -0.56      1.00  0.36     -0.18
## N_Den         0.41 -0.41      0.36  1.00     -0.37
## Phon_Prob    -0.03  0.31     -0.18 -0.37      1.00
```

The correlations can be visualized by plotting them with corrplot. This can be altered to use a specific correlation test using method="" as done above.

```
#install.packages("corrplot", dep=TRUE)
library(corrplot) #Create correlation plot
correlations1 <- cor(data1[,3:7])
corrplot(correlations1, method="circle")
```



Homoscedasticity

Finally, we will test homoscedasticity or the homogeneity of variance using the "lmtest" package. That is, this tests whether the noise or error is the same across all variables. If the

results are significant, we fail to reject the null hypothesis that the data is homoscedastic, otherwise it is heteroscedastic.

```
#install.packages("lmtest", dep=TRUE)
library(lmtest)
lmtest::bptest(lmMod1) #Test for homoscedasticity

##
##  studentized Breusch-Pagan test
##
## data:  lmMod1
## BP = 24.957, df = 5, p-value = 0.000142
```

Multivariate Adaptive Regression Splines

Once the data has been thoroughly explored, an informed decision can be made about an appropriate model for the data. Multivariate adaptive regression splines (MARS) is an adaptable and robust model that is a strong choice in many cases. It does not rely on many of the assumptions of other models such as homoscedasticity, normally distributed data, or low multicollinearity.

The earth command within the "earth" package is used for creating a MARS model. The pmethod argument controls the method for pruning and the options are "none", "backward", "forward", "exhaustive", "seqrep", and "cv". The penalty argument is the generalized cross validation penalty per knot. If penalty=0 is used, only terms will be penalized, not knots. More details for model options can be found at https://cran.r-project.org/web/packages/earth/earth.pdf.

```
#install.packages("earth", dep=TRUE)
library(earth) #MARS
mars1 <- earth(Decon ~ SUBTLwf+AoA+Conc_Mean+N_Den+Phon_Prob, data=data1,pmet
hod = "exhaustive", penalty=1);mars1

## Selected 8 of 16 terms, and 5 of 5 predictors (pmethod="exhaustive")
## Termination condition: Reached nk 21
## Importance: AoA, Conc_Mean, N_Den, SUBTLwf, Phon_Prob
## Number of terms at each degree of interaction: 1 7 (additive model)
## GCV 0.0157487    RSS 1.904409    GRSq 0.8117328    RSq 0.8385457
```

Once the model is created, summary() will display the model, generalized cross validation, residual sum of squares, and the R square value. The evimp() command will show the variable importance and the number of nodes each variable occurs in during model creation, generalized cross validation relative to the most important term, and RSS relative to the most important variable. Root mean square error, mean square error, mean absolute error, and R square values are calculated using the script below.

```
#install.packages("Metrics", dep=TRUE)
library(Metrics) #Fit Metrics
summary(mars1)
```

```
## Call: earth(formula=Decon~SUBTLwf+AoA+Conc_Mean+N_Den+Phon_Prob, data=data
1,
##             pmethod="exhaustive", penalty=1)
##
##                     coefficients
## (Intercept)           0.90511303
## h(32.22-SUBTLwf)     -0.00270751
## h(AoA-5.37)          -0.27390410
## h(AoA-7.81)           0.41768409
## h(AoA-8.45)          -0.17155726
## h(Conc_Mean-3)        0.06923017
## h(126.04-N_Den)      -0.00092140
## h(0.0791-Phon_Prob)  -3.04793995
##
## Selected 8 of 16 terms, and 5 of 5 predictors (pmethod="exhaustive")
## Termination condition: Reached nk 21
## Importance: AoA, Conc_Mean, N_Den, SUBTLwf, Phon_Prob
## Number of terms at each degree of interaction: 1 7 (additive model)
## GCV 0.0157487    RSS 1.904409    GRSq 0.8117328    RSq 0.8385457
```

```
evimp(mars1)
```

```
##           nsubsets   gcv    rss
## AoA              7 100.0  100.0
## Conc_Mean        5  22.1   26.0
## N_Den            4  16.0   20.3
## SUBTLwf          2   8.1   12.1
## Phon_Prob        1   5.5    8.3
```

```
rmse(data1$Decon,as.vector(mars1$fitted.values))
```

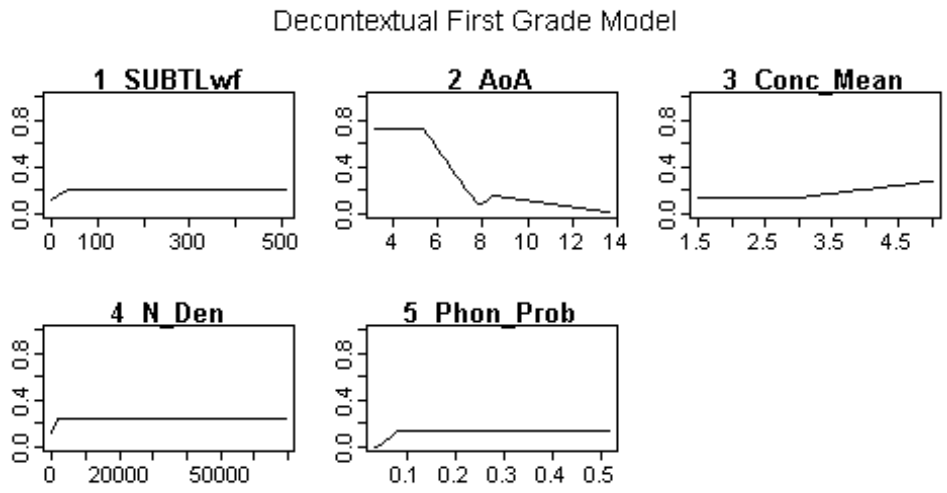
```
## [1] 0.1154017
```

```
mse(data1$Decon,as.vector(mars1$fitted.values))
```

```
## [1] 0.01331754

mae(data1$Decon,as.vector(mars1$fitted.values))

## [1] 0.08576472

pred1 <- predict(mars1, newdata=data1)

rss <- sum((pred1 - data1$Decon) ^ 2)  ## residual sum of squares
tss <- sum((data1$Decon - mean(data1$Decon)) ^ 2)  ## total sum of squares
rsq <- 1 - rss/tss;rsq

## [1] 0.8385457
```

To visualize the model, plotmo() will display the variable plots with all other variables held constant. It is valuable to determine the impact of each variable on word learning. The ylim argument sets the scale of the y-axis so all individual plots are similar. If degree2=FALSE is included, no interaction plots will be included. Only variables chosen by the model will be included in the plot, but all1=TRUE may be included to force all plots to be displayed.

```
#install.packages("plotmo", dep=TRUE)
library(plotmo)
plotmo(mars1, ylim=c(0,1), caption="Decontextual First Grade Model")

##  plotmo grid:    SUBTLwf  AoA Conc_Mean N_Den Phon_Prob
##                      6.9 9.06      2.76  6.49    0.2058
```



The plotmo graphs will be ordered based on the order of the variables during model creation, not variable importance. On each graph, a change in the direction indicates the location of a hinge, which corresponds to hinges in the model.