THE EFFECT OF SPEECH ELICITATION METHOD ON SECOND LANGUAGE PHONEMIC ACCURACY

by

NICOLE HAMMOND CARRASQUEL

B.A. James Madison University, 2001
M.A. University of Central Florida, 2006

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the College of Education and Human Performance
at the University of Central Florida
Orlando, Florida

Summer Term
2017

Major Professors: Marcella A. Farina and Keith S. Folse

© 2017 Nicole Hammond Carrasquel

ABSTRACT

The present study, a *One-Group Posttest-Only Repeated-Measures Design*, examined the effect of speech elicitation method on second language (L2) phonemic accuracy of high functional load initial phonemes found in frequently occurring nouns in American English. This effect was further analyzed by including the variable of first language (L1) to determine if L1 moderated any effects found. The data consisted of audio recordings of 61 adult English learners (ELs) enrolled in English for Academic Purposes (EAP) courses at a large, public, post-secondary institution in the United States. Phonemic accuracy was judged by two independent raters as either approximating a standard American English (SAE) pronunciation of the intended phoneme or not, thus a dichotomous scale, and scores were assigned to each participant in terms of the three speech elicitation methods of word reading, word repetition, and picture naming.

Results from a repeated measures ANOVA test revealed a statistically significant difference in phonemic accuracy ($F(1.47, 87.93) = 25.94$, $p = .000$) based on speech elicitation method, while the two-factor mixed design ANOVA test indicated no statistically significant differences for the moderator variable of native language. However, post-hoc analyses revealed that mean scores of picture naming tasks differed significantly from the other two elicitation methods of word reading and word repetition.

Moreover, the results of this study should heighten attention to the role that various speech elicitation methods, or input modalities, might play on L2 productive accuracy. Implications for practical application suggest that caution should be used when utilizing pictures to elicit specific vocabulary words–even high-frequency words–as they might result in erroneous productions or no utterance at all. These methods could inform pronunciation instructors about

best teaching practices when pronunciation accuracy is the objective. Finally, the impact of L1 on L2 pronunciation accuracy might not be as important as once thought.

ACKNOWLEDGMENTS

The University of Central Florida has built my professional life, starting in 2004 when I first entered the doors of Dr. Keith Folse’s office on the fifth floor of Colbourn Hall to learn more about the difference between a master’s degree in TESOL and a graduate certificate. At this meeting, he encouraged me to go for the masters as it would open more doors for me in the future job market wherever that might lead me. I took his advice, prompted greatly by his recommendation for me to receive an assistantship at the Center for Multilingual Multicultural Studies (CMMS), the intensive English language program on campus at that time. When the news of the assistantship came to me, it solidified my decision to choose UCF over other colleges I had applied to at that time. It also allowed me to work at one of the most important jobs of my life, a job where I grew professionally and personally. My sincerest gratitude goes to Dr. Folse for opening these doors for me at UCF.

CMMS was a small haven of fun and energetic teachers who had a passion for teaching English to internationals from all over the world. What started out as a part-time assistantship developed into a full-time adjunct position upon graduating with my masters in TESOL, and four years later, I was offered a full-time instructor position there. During my 11 years of working there, I was afforded the opportunity to develop professionally by attending and presenting at various TESOL conferences, including TESOL International. Supported greatly by the CMMS director at the time, Myrna Creasman, I was able to develop several teacher training programs and even go overseas in this capacity. Myrna entrusted me with a lot of responsibilities over my years of working at the center, and in doing so, my resume just kept getting stronger. Moreover, I learned about the type of teacher I wanted to be. Myrna encouraged not only hard work but

also making the job fun. She supported me when I decided to run a drama club at the center, something that enriched my life and made my job one of the best I have ever had. As a result, I credit much of this endeavor to Myrna Creasman and her amazing CMMS.

I am also grateful for the wonderful teachers, staff, and students I met during my time at CMMS. They became my dear friends and supporters of my eventual decision to further my education with the PhD. In particular, there are two people who deserve mention here who were very instrumental in supporting me. The first person is Dr. Marcella Farina, my mentor and colleague of over 10 years and a great friend since the first day I walked into the doors of CMMS. She has been an invaluable support to me during all my time at UCF. Marcella gives more than one hundred percent of her time to help people and definitely gave me more time these last few years as the chair of my dissertation committee. I cannot speak for other chairs, but I can attest to the many hours that she spent with me, both in person and through the very detailed feedback she always provided me. She is an excellent example of a model educator and has always helped me to be the best teacher (and student) I could ever hope to be. To her, I give the most credit for accomplishing this goal in the time that I was able to do so.

The second person I give credit to is Katya Goussakova, a fellow CMMS instructor at the time I first started working there and now one of my best friends. She is the person who is largely responsible for why I got myself into this sometimes-grueling task of getting my PhD, and she has always supported me to better myself professionally. When we met for coffee one fateful spring day, she informed me of her decision to start the PhD program at UCF, encouraging me to do so with her. Considering we both had young children about the same age and would go through all of it together made the decision easier. Since then we have celebrated

our victories, commiserated our setbacks, and identified with each other's struggles both as students, mothers, wives, and professionals. She has had my back during any challenges that arose, and there were many. Truly this experience would not have been the same without her.

There are also some friends who deserve credit for my survival over these years. To my entire PhD cohort, but in particular to Alex Davies and Laura Monroe, whose frequent visits with me to Bar Louie and other restaurants ended up being the best planning and support sessions I could have asked for during our PhD course work time. And I could never forget to thank my best friend "Loco," known to others as Raúl Fernández Sumano, and his soon-to-be famous stress-management sidekick "Furball," who made my whole experience in the PhD program more manageable and fun. Frequent coffee breaks, helpful discussions, inventive stories, and good laughs were the only way I survived some of the more difficult moments over the years. Without his support and sense of humor, I might not have enjoyed this journey as much as I did. To you, my friends, I give you credit for relieving my stress and for all the fun I had.

To my committee members, this degree is in your footsteps. My vice-chair Dr. Keith Folse has been an amazing mentor and unofficial advisor to me through both my masters and PhD. His mark on the field of TESOL is truly inspiring. Dr. Kerry Purmensky stepped in when she was most needed, and she has always been a wonderful colleague and supporter throughout the years. She supported my research efforts from the very beginning of my PhD experience and was instrumental in securing my target population for this study. To Dr. M. H. Clark I give credit for helping me better understand the statistical and research method concepts from when I first started in the PhD. She brought to light many important distinctions for me and offered very thorough feedback. She was always willing to meet with me, sometimes for hours, to go through my data analysis and fill in some gaps in my learning. I thank her for her much-needed expertise here. I am so happy to have each and every one of these members on my committee.

Quite a few other people deserve credit in helping my dissertation study come to fruition, including Dr. Olga Bedoya, the academic director of Global UCF's Achievement Academy who supported my research and allowed me to conduct both a pilot study and my dissertation study with the students enrolled in her program; Miranda, the computer tech support person without whose help I might not have gotten nearly as many useful audio files during my data collection; Aimee Schoonmaker and Christina Torres, my fellow instructors, who were so supportive and flexible in allowing me to conduct this study in their classes; Dr. Lihua Xu for her statistical guidance in the CASTLE lab; and all of my professors who instructed me and helped me to form my research questions and design, including Dr. Glenn Lambie, Dr. Debbie Hahs-Vaughn, Dr. E. Lea Witta, Dr. Stephen Sivo, Dr. David Boote, Dr. Haiyan Bai, Dr. Florin Mihai, Dr. Gergana Vitanova, Dr. Linda Rosa-Lugo, Dr. Malcolm Butler, Dr. Michele Regalla, Dr. Barbara Ehren, Dr. Elsie Olan, and Dr. Joyce Nutta, our respected TESOL PhD coordinator.

Finally, it goes without saying that in order to do the work required of a PhD with two young children, the only way is with family support. My supportive husband Alejandro took on so many extra responsibilities and helped make it possible for me to work late nights and sometimes on the weekends. It was, to be sure, not an easy experience for us, but he made it possible. I know he will be glad when it is finished, but he also kept a good attitude throughout, and I recognize that it would not have been possible to do without him supporting me at home. To all these people (and others working behind the scenes), this culmination of my academic success goes to you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AoA	Age of Arrival
ANOVA	Analysis of Variance
COCA	Corpus of Contemporary American English
CPH	Critical Period Hypothesis
EAP	English for Academic Purposes
EL	English Learner
ESL	English as a Second Language
IELTS	International English Language Testing System
L1	First Language
L2	Second Language
LOR	Length of Residence
NS	Native Speaker
NNS	Nonnative Speaker
POEC-S	Proficiency in Oral English Communication Screening
SAE	Standard American English
SEMi	Speech Elicitation Method Instrument
SLA	Second Language Acquisition
TOEFL	Test of English as a Foreign Language

CHAPTER ONE: INTRODUCTION

Inquiry into second language acquisition (SLA) has looked at the various factors that influence this process. From maturational factors inherent to all language learners to sociocultural factors, learning another language is multifaceted and has shown great variety both across and within language learners. With a myriad of areas involved in the process of acquiring another language, perhaps no other area is more vital to overall language competency than pronunciation. For decades, pronunciation experts have stressed that comprehensibility and intelligibility need to be the focus of pronunciation pedagogy (Derwing & Munro, 2009) rather than attempting to reduce accentedness and encourage nonnative speakers (NNSs) to perform like native speakers (NSs). However, nonnative accentedness can both impede communication, and affect access to certain groups, speech communities, and opportunities (Pavlenko & Blackledge, 2004; Norton, 2013). The truth is that having a foreign accent, in many circles, has a negative connotation and, in fact, has been shown to influence rater judgments of NNS oral proficiency (Derwing & Munro, 2005; Kang, Rubin, & Pickering, 2010; Rajadurai, 2007).

Nonnative accentedness is derived from both differences in the production of phonemic segments as well as prosodic features (Kang, Rubin, & Pickering, 2010). Advocates of a more global approach to instruction (Celce-Murcia, Brinton, & Goodwin, 1996; Firth, 1992; Pennington & Richards, 1986) have argued that prosodic features, such as speech rate, stress, intonation, pitch, and rhythm, impact intelligibility more than segmental drills commonly used in pronunciation lessons (Derwing & Munro, 1997; Derwing & Rossiter, 2003). Many in the field of SLA concur that there are a number of factors contributing to pronunciation accuracy regardless of whether the focus is on comprehensibility, intelligibility, or accentedness (Munro &

Derwing, 1994, Piske, MacKay, & Flege, 2001), yet what constitutes accuracy varies greatly among studies. Furthermore, factors peripheral to phonological features, including listener attitudes, can get muddled into the discernibility of intelligible language (Lindemann, 2010).

Turning to the field of Teaching English to Speakers of Other Languages (TESOL), there are many English accent varieties used around the world. Kashru claims that with the number of NNSs of English outnumbering the number of NSs of English (1997, p. 214), caution should be used when teaching English pronunciation to a “native-speaking model” as it needs to be considered in light of these many NS and NNS accent varieties (Jenkins, 2002). What constitutes “standard” English is negotiable and depends greatly on the area and context in which it is used, so accuracy is often defined in terms of the local variety in which the speech is evaluated (Kang & Moran, 2014).

The pendulum continues to swing over whether to focus pronunciation instruction on accuracy or intelligibility, and under more recent trends, intelligibility is winning (Derwing & Munro, 2009). In response to the more traditional grammar translation and audiolingual approaches espoused in past language learning, communicative language teaching approaches still popular in today’s language classrooms stress fluency and meaning over accuracy. Nowadays a more informed direct approach, one that focuses on meaning but which also involves awareness raising, is gaining in popularity to better balance communicative language learning (Grant, 2014).

Background of the Problem

Even though there is a trend to empower international varieties of English and other languages (Crystal, 1997; Jenkins, 2002), positive and negative attitudes about specific dialects and accent varieties still prevail. Variation among these attitudes has been attributed to a number of factors, including the degree of accentedness (Brennan & Brennan, 1981), familiarity with accents (Chiba, Matsuura, & Yamamoto, 1995; Dalton-Puffer, Kaltenboeck, & Smit, 1997); speaker traits, primarily those of status and solidarity (Bayard & Green, 1995; Chong & Tan, 2013; Giles, 1970), knowledge of speaker nationality (Yook & Lindemann, 2013), ethnic group affiliation (Gatbonton, Trofimovich, & Magid, 2005), and identity categories (Bresnahan, Ohashi, Nebashi, Liu, & Shearman, 2002). Regardless of the comprehensibility of each variety of English, these attitudes arguably impact the use of certain dialects or accents, and people desire to emulate certain accents or diminish their own accent in any given language. More specifically, both NSs and NNSs of English have shown a preference for native forms over nonnative forms (Scales, Wennerstrom, Richard, & Wu, 2006; Timmis, 2002).

Measurements of L2 phonemic accuracy become important under this view, yet determining what constitutes “accurate” can be problematic and ultimately depends on the setting of each investigation. As Kang and Moran (2014) added, variability exists even within NS varieties of English (e.g. British English, New Zealand English, American English). Typical measurements of L2 phonemic accuracy are evaluated in dichotomous ways (Liu & Fu, 2011; Riney & Flege, 1998); that is, the production is either accurate or inaccurate. However, inaccurate productions should be understood as a deviation from a specific NS norm and not necessarily as erroneous or incorrect because the inaccurate phoneme could actually be a feature of some nonstandard varieties.

Background to Corpus Linguistics and the Functional Load

To more accurately depict true language use in a specific setting, the relatively new field of corpus linguistics has shed light on the many perceptions, and misperceptions, that exist about how language is used (Reppen, Fitzmaurice, & Biber, 2002). Biber and Conrad (2001) argue that prior to corpus linguistics, most knowledge about language use was based on linguists' observations and studies. Defining a corpus as a "large, representative database of spoken or written texts," they also discuss how corpus linguistics can help to identify the most salient linguistic features for textbook publishers to focus on and language teachers to teach. The first corpus was compiled in 1962 (the Brown Corpus; Kučera & Francis, 1967), and since then corpora have been created for quite a few purposes, including compiling dictionaries and informing research efforts. Nowadays, the use of corpora has steadily increased with computer-assisted analyses being conducted on corpora containing millions of words to offer a fast way to study language patterns and natural examples of language use (Biber & Conrad, 2001; Biber, Conrad, & Reppen, 1998).

Among the many often-cited corpora is the Cambridge and Nottingham Corpus of Discourse in English (CANCODE), part of the larger Cambridge International Corpus, which includes millions of written texts and spoken conversations transcribed for both British and American English. Focused more specifically on American English, several other corpora have been created and add to the research investigating this English variety. Among these is the Corpus of Contemporary American English (COCA), which contains over 520 million words and subsequent word frequency lists derived from the corpus (Davies, 2015). A smaller corpus of English with a focus on English spoken in a university setting is the Michigan Corpus of

Academic Spoken English (MICASE), compiled between 1997 and 2001 and containing 197 hours of recorded speech ranging from academic lectures, one-on-one office hour interactions, and small group discussions (Simpson, Briggs, Ovens, & Swales, 2002).

When looking at frequencies of phonemes, the functional load (Brown, 1991; Catford, 1987) can serve as a way to prioritize those segmental features that are more likely to influence listeners' comprehension. In essence, functional load theory suggests that phonemes with a relatively low functional load distinguish fewer words in spoken English and, therefore, are less likely to be as problematic as phonemes with a relatively high functional load. It is important that SLA research studies consult various corpora and/or the functional load to ensure more meaningful and relevant outcomes.

Background to Speech Elicitation Methods

Linguistic variation studies have shed light on pronunciation differences that ELs have and sometimes willingly employ in their English productions. L2 variation research stems from sociolinguistics (Labov, 1966) with one goal being to identity internal and external factors that predict the use of variant forms. The term *interlanguage* (Selinker, 1972) views L2 learning as developing in a systematic and dynamic way with research on L2 variation depending on the linguistic and social contexts of the speech event. In addition, language learners respond to various speaking tasks in different ways, and their social identities can impact their L2 language use and phonological attainment in that language (Norton Pierce, 1995; Norton, 2000). Related to this area of sociolinguistic inquiry is the influence of speech settings or topics on phonological acquisition and accuracy. The way in which speech is elicited can play a vital role in its ultimate success or failure, however that may be defined.

Speech elicitation may influence language production, so task type and content are factors that should be investigated in any study of L2 pronunciation ability. Two of the most common ways to elicit speech are via auditory prompts or visual prompts, so typical elicitation methods involve repetition, response, reading, or pictures. Studies incorporating these elicitation methods (Beebe, 1980; Borden, Gerber, & Milsark, 1983; Derwing, Munro, & Wiebe, 1998; Elliott, 1997; Liu & Fu, 2011) often have students repeat words, repeat longer phrases or sentences, read isolated word lists or sentences, and/or respond to either pictures, written prompts, or interview questions.

Research has indicated that the ability to mimic can aid in obtaining more native-like accent ratings (Flege, Yeni-Komshian, & Liu, 1999; Purcell and Suter (1980), so the use of repetition as an elicitation method is a valuable option. Moreover, the literature on word reading tasks has shown some facilitative and some non-facilitative effects as to the role of orthography on L2 pronunciation accuracy (Frith, Wimmer, & Landerl, 1998; Koda, 2016; Landerl, 2000; Wang & Koda, 2007), so reading prompts are also viable elicitation methods to investigate. Finally, picture naming has been investigated in the TESOL field but primarily in bilingual studies (Ivanova & Costa, 2008; Kroll, Bobb, & Wodniecka, 2006; Sholl, Sankaranarayanan, & Kroll, 1995), yet the field of speech and language pathology has examined this elicitation method as a valid way of prompting specific words and phonemes for determination of speech-sound disorders (Morrison & Shriberg, 1992; Wolk & Meisler, 1998).

It is important to note that these studies including elicitation method as a variable examined different L1s and L2s, as well as different age groups. In addition, some studies looked at pronunciation accuracy of real words while others looked at this phenomenon within nonwords. Considering that children learn differently from adults and that languages operate under differing degrees of orthographic depth; it is critical that all of these factors be taken into consideration when investigating the effect of speech elicitation methods on pronunciation accuracy.

Statement of the Problem

Derwing and Munro (2009) noted the following trend in SLA: “the decline of audiolingualism led to a concomitant marginalization of pronunciation research and teaching” (p. 476). Grant (2014) also discussed this change of attention toward pronunciation over the course of various popular language teaching approaches. During the communicative language teaching era, there was a backlash against traditional approaches to L2 pronunciation instruction and assessments which tended to focus primarily on segmental features, minimal pair drills, and decontextualized material. The focus of English as a second language (ESL) classrooms shifted from accuracy, often based on a NS norm, to meaning and intelligibility. This trend of abandoning the goal of perfect, native-like pronunciation continues to the present day, and while L2 pronunciation research has steadily been increasing, it is still lacking compared to other areas of inquiry in SLA (Grant, 2014). Furthermore, many L2 instructors have indicated feeling unprepared, and even fearful, to teach and assess NNS pronunciation (Breitkreutz, Derwing, & Rossiter, 2002, MacDonald, 2002). Surveys conducted by Murphy (1991) revealed that only approximately half of all MA in TESOL programs offer specific courses on phonology, adding to this general lack of understanding of this important area in SLA.

Despite current trends to encourage international and other nonstandard varieties of English, many language learners still strive to sound more native-like in their L2 pronunciation

(Chiba, Matsuura, & Yamamoto, 1995; Dalton-Puffer, Kaltenboeck, & Smit, 1997; Scales, Wennerstrom, Richard, & Wu, 2006; Timmis, 2002). In other words, simply being comprehensible and intelligible in their pronunciation is not enough for many ELs; they also hope to reduce their accentedness and learn more standard forms of the language, forms that will allow them access to speech communities and imagined identities (Kanno & Norton, 2003; Pavlenko & Norton, 2007).

Finally, often paired with speaking skills, pronunciation assessments tend to focus on accuracy in production, while the role of input or exposure to the language has largely gone unnoticed (Flege, 2012). Many adults learn another language after already achieving literacy in their L1. So prevalent is the written word in pronunciation training and practice that educators and adult language learners alike rely on graphical representations of the language to assist in L2 pronunciation efforts. The International Phonetic Alphabet (IPA) is just one example of this trend. Although quite a few studies have examined the effects of orthography on reading acquisition, few have looked specifically at how orthography influences speech, and in particular, phonemic accuracy. Even fewer studies have investigated auditory elicitation methods and their impact on L2 phonemic accuracy, and research on picture naming as a potential method of eliciting L2 pronunciation is even scarcer. All in all, the various ways in which language learners are exposed to L2 pronunciation activities and assessments might play a role in their productive accuracy, yet they are often overlooked.

Rationale of the Study

The relative lack of attention to and understanding of L2 pronunciation (Grant, 2014) can cause language teachers to fend for themselves when it comes to finding support for teaching pronunciation. Stand-alone pronunciation classes are rarely offered in L2 teaching facilities (Munro & Derwing, 2006), and pronunciation is instead infused throughout the program, typically incorporated marginally into speaking classes. Moreover, some L2 teachers are even fearful of teaching pronunciation, prompted by feelings of unpreparedness to teach in this SLA area (Breitkreutz, Derwing, & Rossiter, 2002, MacDonald, 2002). Because of the relative scarcity of L2 pronunciation research and lack confidence in teaching pronunciation, there is a need for further investigation into this important yet undermined area of SLA.

In addition, ELs' preferences are contributing to this need for more focused pronunciation assessments that evaluate accuracy according to SAE norms (Scales, Wennerstrom, Richard, & Wu, 2006; Timmis, 2002). Pronunciation intelligibility, while important, is just not enough for many language learners. At the same time, decontextualized minimal pair word lists to practice segmental accuracy are also insufficient. Corpus linguistics has offered valuable information to SLA research with regard to prioritizing salient features of the language to practice and assess. According to Kang and Moran (2014), "reflecting the hierarchy of phonemic errors, segmental features are often discussed in terms of functional loads" (p. 177). What appears to warrant further research, therefore, is a more meaningful way to investigate L2 phonemic accuracy, one that is corpus-informed and includes the most prevalent words in American English and one that considers the functional load.

Finally, because of a relative paucity of SLA studies investigating speech elicitation methods and their impact specifically on L2 phonemic accuracy, input modalities, or ways in which L2 learners are exposed to the language, need further investigation. The importance of input on L2 pronunciation is just recently gaining ground (Flege, 2012), so an investigation of various speech elicitation methods and their effect on L2 phonemic accuracy is in order. In sum, the three elicitation methods of word reading, word repetition, and picture naming have not been examined together for their impact on L2 phonemic accuracy, so the present study will attempt to fill this void.

Research Questions

Based on the need for further inquiry into L2 pronunciation, a desire to measure L2 phonemic accuracy in a meaningful way, and the potential influence that speech elicitation methods might have on this important SLA area, the present investigation aimed to answer the following research questions:

1. Is there a statistically significant difference in L2 phonemic accuracy based on the speech elicitation methods of word reading, word repetition, and picture naming of high functional load initial phonemes found in high-frequency nouns occurring in American English?
2. Does native language moderate the effect of speech elicitation method on L2 phonemic accuracy of high functional load initial phonemes found in high-frequency nouns occurring in American English?

Research Hypotheses

In response to Research Question 1, the null hypothesis was that there would be no difference in phonemic accuracy scores based on speech elicitation method. However, it was hypothesized that there would be a statistically significant difference ($p < .05$) in L2 phonemic accuracy based on the speech elicitation methods of word reading, word repetition, and picture naming of high functional load initial phonemes found in high-frequency nouns occurring in

American English. With regard to Research Question 2, the null hypothesis was that native languages would not moderate the effect of speech elicitation method on phonemic accuracy scores. It was hypothesized, however, that native language would moderate the effect of speech elicitation method on L2 phonemic accuracy of high functional load initial phonemes found in high-frequency nouns occurring in American English at an alpha level of less than .05.

Definition of Terms

Several terms are important to define as they relate to the research questions. The first term is *phonemic accuracy*. Phonemic accuracy, the dependent variable in the present study, was judged by NS raters as either approximating or not approximating the SAE pronunciation of phonemes. Phonemic accuracy is based on Derwing and Munro's (2009) operationalization of *accent*, which was examined in terms of listener perception of speech. According to Derwing and Munro, both native and nonnative listeners have shown high reliability ratings (with Pearson correlations of $p \geq .9$) for detecting accents. They further defined three dimensions of accent: salience, comprehensibility, and intelligibility. *Salience,* also referred to as *accentedness*, is "how different a pattern of speech sounds compared to the local variety" (p. 478). *Comprehensibility* is "the listener's perception of how easy or difficult it is to understand a given speech sample" (p. 478). *Intelligibility* is "the degree of a listener's actual comprehension of an utterance" (p. 479). For the purposes of the present investigation, only accentedness based on the standard American variety was interpreted. Certain phonemes, such as the use of the trill /r/, while acceptable by other accent varieties, were not counted as approximating SAE and, therefore, inaccurate. It is important to understand that accuracy not be confused with correct but rather in terms of approximating just one standard English variety.

The second term, *speech elicitation method*, was defined as the way in which speech, and for the purposes of this study, specific phonemes, were elicited. There are many ways to elicit word responses from participants, but for the purposes of this study, the following three were included: 1) *word reading* represented by orthographic depictions of the instrument vocabulary presented in sanserif font on a beige background, 2) *word repetition* consisting of audio recorded spoken utterances of the instrument words, and 3) *picture naming* consisting of colored, two-dimensional images representing the target words of the instrument.

The *functional load* was first coined by Jakobson (1931) and Mathesius (1931) and further developed by Hockett (1967) to analyze the phonemes inside the phonological system of any language; more specifically, it refers to the amount of work that a phoneme does to distinguish words in communication. High functional load phonemes occur more frequently and in more contrasts than low functional load phonemes. For instance, in English there are hundreds of word pairs that differ only in /p/ and /b/, such as *pack* and *back* or *cup* or *cub*; on the other hand, there are relatively few examples of word pairs that differ in /θ/and /t/, such as *thank* and *tank* or *bath* and *bat.* The first contrastive examples involve the phonemes in initial positions, while the second examples illustrate final position contrasts. Functional load was introduced into the field of SLA by both Brown (1988) and Catford (1987), the latter who quantified these phonemic contrasts in initial and final positions with percentages based on the number of times they occurred per thousand words of text. Catford listed initial consonant, final consonant, and vowel contrasts in order by their relative functional load with the highest pair represented by 100%. For the purposes of the present study, high functional load initial phonemes were defined as possessing frequency percentages of more than 50%, based on Catford's relative functional load (p. 89-90).

Phonemes were defined as the phonological realization of the graphemes represented by the English alphabet. English speakers typically recognize 26 letters in the English alphabet; however, many may not know there are approximately 44 phonemes, or sounds, in spoken American English. In the present study, phonemes were categorized by voicing, place of articulation, and manner of articulation and were written using the IPA symbols. Only those phonemes classified as possessing high relative functional loads were examined.

High-frequency nouns were defined in terms of what corpus findings have revealed to be the most frequently occurring in American English. Because this study was administered in an American English setting, the COCA (Davies, 2015) was consulted, which offers free word frequency lists, one of which includes the 5,000 most frequently occurring words in American English. Thus, only the most frequently occurring nouns in American English from the COCA list were included in the present investigation.

Finally, as the second research question investigates native language, *native language* was defined as the L1 of a person. The question of which English phonemes are more difficult for ELs to pronounce can vary based on the native language of the speaker. While not all segmental issues stem from the L1, some error patterns frequently occur based on L1 (Swan & Smith, 2001) so including native language as a variable is worthwhile.

Contributions of the Study

This study contributed to the body of research on input modalities in SLA, which is lacking (Flege, 2012). More specifically, it examined the three speech elicitation methods of word reading, word repetition, and picture naming. The current study also added more understanding to the body of research on L2 pronunciation, an area which is relatively misunderstood and lacking in research compared to other SLA areas (Grant, 2014). In essence, it informed L2 researchers, educators, and students of the importance of considering input modalities in both instruction and practice of pronunciation accuracy. While this study did not find one more optimal way of achieving more “native-like” accuracies for its participants, it discovered one less reliable way, that of using pictures to elicit specific sounds.

The second contribution is related to the instrument created for this study. If later validated, the instrument is flexible in that it offers three elicitation methods whereby a teacher could extract a method most suitable to a particular population. It could be used generally for pronunciation placement or customized to measure specific phonemes based on instructional components, thus offering an efficient assessment for placement, diagnostic, proficiency, and progress within a language program. At the very least, it could be used as a template of an instrument creation process for researchers of other languages. The process by which the instrument for the present study was constructed utilized detailed steps and pulled from a corpus to have a more meaningful list of items. Such an instrument-creation process could be replicated or revised to assess other English learners or other phonemes.

Thirdly, the results of this study showed that there are a variety of ways in which speech for phonemic accuracy can be gathered and assessed. This knowledge can contribute to language teacher readiness in teaching pronunciation. Better informed teachers who have greater knowledge in this specialized area of SLA can facilitate their students’ L2 pronunciation goals. Having a better understanding of ways to elicit speech for phonemic accuracy can provide L2 teachers with the confidence they might lack. In the end, the findings from this study indicated that there are better and worse ways to elicit words for pronunciation accuracy, so teachers who are better prepared, more knowledgeable, and more confident language teachers are critical to the success of language learners.

The final contribution is to the language learners themselves. Individual learner differences exist, and language learners operate under varying learning styles. Because of this study, L2 learners might be more cognizant of their own learning tendencies and which outcomes to manipulate for their own language learning gains when it comes to pronunciation accuracy. They could then be able to seek help and be better informed about ways to sound more native-like, if they so desire to speak with this variety.

CHAPTER TWO: LITERATURE REVIEW

In order to review the literature related to eliciting L2 speech for the purpose of evaluating phonemic accuracy, it is necessary to identify relevant theoretical frameworks that provide background knowledge and a rationale for the proposed investigation. To this end, the areas to be reviewed are the following: factors influencing L2 phonological acquisition, the functional load, speech elicitation methods, and measurements of L2 phonemic accuracy. First, many factors have shown to influence phonological acquisition, but several interspeaker and intraspeaker factors pertaining to the present investigation are addressed. In addition, the functional load has focused on frequency in phonemes and serves as the core element in several pronunciation studies' instruments. Next, speech elicitation methods are examined more thoroughly as one of the factors influencing phonemic accuracy. Finally, measurements of L2 phonemic accuracy are investigated, broken down into two L2 pronunciation principles that have influenced how data has been gathered, what tasks have been included, and how data has been interpreted. By understanding what other research has shown, the foundation for future research can be established and expanded. It is the hope that another brick might be added to the knowledge foundation of what contributes to phonemic accuracy and ultimate L2 phonological attainment for adult language learners.

Factors Influencing Second Language Phonological Acquisition

The Critical Period Hypothesis (CPH) is the notion that there is a window of time in order for successful language acquisition to occur, after which language learning becomes impossible or labored, preventing an ultimate level of attainment. Compelling evidence exists for a critical period in L1 acquisition with examples of feral children who were not exposed to any human language input or output until later in their childhood or adolescence. Convincing cases, such as that of Genie (Curtiss, 1977), a girl who was neglected by her parents and brought up without hardly any human contact until the age of 13, seem to prove the existence of a critical period, at least in terms of L1 acquisition. Genie had great difficulty in successfully acquiring an L1 despite hard work with linguists, personal effort, and even high motivation to learn.

Critical Period Hypothesis for SLA

The question of whether a critical period exists for SLA has also been investigated yet with less conclusive evidence. Lenneberg (1967) observed that unlike adults, children were able to learn native-like language without effort via mere exposure. He hypothesized that the brain capacity for language learning decreases with age and that post-pubertal language acquisition would, therefore, not be as automatic as pre-pubertal language acquisition. Twenty years later, more research supported these claims with Bornstein (1987) outlining five stages to the biological critical period: (1) organismic system, (2) environmental input, (3) onset, (4) duration, and (5) terminus. In L1 acquisition, answers to the last three stages are recognized with birth being the onset, an approximate length of learning and instruction at around five to seven years, and a terminus of around 12 years. However, Bornstein also claimed that there is too much variability in SLA to provide reliable answers for the exact onset, duration of learning and instruction, and a time after which it is no longer possible to learn.

The distinction between adult and child learning processes is also addressed in Bley-Vroman's (1989) influential Fundamental Difference Hypothesis, which claimed a crucial difference between children learning their L1 and adults learning their L2 and linked the deficit to a lack of access to Universal Grammar (UG) in adult learning. Birdsong (1999) defined the

CPH by stating that, after a certain age, L2 learners experience increased sensitivity in acquiring the language due to a loss of neural plasticity whereby language learning generally plateaus and gradually declines. All in all, a critical or sensitive period for SLA, although extensively hypothesized over the last decades, has not yet been empirically proven.

One of the aspects of SLA that has shown some empirical evidence in favor of a critical period is that of pronunciation, or foreign accent, in the target language. According to Scovel (1988), speech is physiological in nature and, thus, pronunciation is the one area influenced by the CPH; in other words, there could be a "selective sensitive period" for the acquisition of phonology. This area of language learning has proven to be one of the most challenging for NNSs, and what appears to be consistent in the research findings is that of phonological acquisition possessing a critical period, if not a sensitive period, after which it becomes much more difficult for adults learning another language to obtain a native-like accent in the target language (Asher & Garcia, 1969; Flege et al, 1999; Oyama, 1976; Piper & Cansin, 1988; Scovel, 1988; Thompson, 1991).

More specifically, L2 research studies have investigated a number of factors that influence L2 phonological acquisition. Among the many factors that have been correlated to degree of foreign accent are age of arrival, time of exposure to the L2, L1 use, instruction, motivation, aptitude for oral mimicry, and degree of concern for pronunciation accuracy (Moyer, 2004). Moyer stated that what is more understood is that L2 phonological attainment remains quite variable, and there are multiple factors wielding an impact on the process. As Moyer succinctly concluded, "mechanisms behind late learner attainment are multi-variate and complex, leading to great variation in outcome" (p. 46). General consensus nowadays is that individual L2 achievement is inconsistent, and some adult L2 learners are able to become indistinguishable from NSs, so a critical period might not exist for SLA (Herschensohn, 2007).

Interspeaker Factors

Many factors are known to impact the phonological acquisition process, some of which are interspeaker factors. In a comprehensive summary of these factors from the past 30 years, Piske, MacKay, and Flege (2001) identified seven major influences examined in relation to L2 phonological acquisition: (a) age of L2 learning, (b) length of residence in the target country, (c) gender, (d) formal instruction, (e) motivation, (f) language learning aptitude, and (g) amount of L1 use. The first two, both interspeaker factors, are the ones most discussed in the research.

Age of L2 learning. Of the interspeaker factors, one of the most highly investigated is that of age of arrival (AoA), or age of onset. Many studies have been conducted showing a positive correlation between AoA and foreign accents. Asher and Garcia (1969) studied 71 Cuban immigrants to the United States between the ages of seven and 19 who had arrived approximately five years before the study and found that as the AoA of their participants increased, so did their foreign accent ratings. Oyama's (1976) study of 60 Italian immigrants to the United States also discovered higher ages of arrival to be correlated to stronger foreign accents ($F(2, 54) = 70.27$, $p < .01$). Similarly, Piper and Cansin (1988) investigated 29 ESL learners living in Canada who had arrived between the ages of six to 28. NS judges rated the participants' accents, and those who had arrived in Canada as children demonstrated higher native-accent ratings than those who had arrived later. Other studies supporting a strong relationship between AoA and L2 accents were Thompson's (1991) study of 36 Russian-born immigrants to the United States, with AoA accounting for 66% of the variance in accent scores, and Flege, Yeni-Komshian, and Liu's (1999) investigation of 240 Korean immigrants to the United States with 13% of the variance in accent scores accounted for by AoA.

Despite these findings, Olson and Samuels (1973) found a facilitative effect of age on pronunciation accuracy in a controlled imitation-based study. In their study, three groups of 20 children, 20 adolescents, and 20 adults with no previous foreign language instruction for German were exposed to repetition drills of German phonemes. After thirteen phonemic instruction sessions on how to model their pronunciation after a German speaker, pre and post scores of their production of German phonemes were compared for accuracy. Findings showed that adults were superior to children in mimicking pronunciation in a foreign accent ($F(2, 53) = 5.17$, $p < .009$). Snow and Hoefnagel-Höhle (1977) also found similar findings in their controlled longitudinal study of 136 native English speakers ranging in age from five to 31. When tested on their ability to imitate five different Dutch words over three time periods, the older participants held an initial advantage over the younger participants ($H = 31.96$, $\chi^2 = 19.68$, $p < .05$); nevertheless, over time this advantage disappeared.

In addition, in two separate experiments, Bongaerts, van Summeren, Planken, and Schils (1997) found that age of onset of the L2 did not necessarily affect late L2 learner ability to obtain native-like accent ratings. In the first experiment, three groups of participants–NS controls, highly successful Dutch learners of English, and Dutch learners of English of various proficiencies–were asked to speak about a personal experience and read aloud a list of sentences and a list of words in English. Their accent was rated by native speaking judges, and results showed that some of the Dutch participants who had learned English after the age of 12 were undistinguishable from the native English speakers. In the second experiment, more participants were added to the three groups, and only the sentence reading task was analyzed. Results again showed that the group of highly successful learners (*M* = 4.61) held similar accent ratings to the NSs (*M* = 4.84) with some of the late learners receiving higher scores than some of the NSs of English. Lastly, while Moyer's (1999) study of 25 nonnative German speakers found age of onset to be a significant factor, other factors like motivation were actually found to be stronger indicators of degree of foreign accent.

Length of residence. Also noteworthy of the interspeaker factors is length of residence (LOR) in the L2 environment. For instance, Asher and Garcia (1969) found that of the population studied (*N* = 71), 51% of the Cuban children between the ages of seven and 19 who had been living in the United States for more than five years had near-native pronunciation ratings, while only 15% of the same children who had been living in the United States for four or fewer years had the same near-native ratings. Furthermore, Purcell and Suter (1980) examined 61 NNSs of English and found length of residence in the United States to be one of the greatest predictors of pronunciation accuracy in English, and it accounted for an additional 11.41% of the variance in scores.

However, the results of other studies have found differing effects of LOR. In Oyama's (1976) study of 60 Italian-born immigrants who had been living in New York between five to 18 years, results showed virtually no effect between degree of accent and the number of years they had been living in the United States. Similar findings came from Piper and Cansin's (1988) study of 29 advanced ESL learners in Canada. They were unable to confirm the hypothesis that a greater number of years in the L2 environment equated to more native-like accents. Riney and Flege (1998) conducted three experiments that examined 11 Japanese learners of English on their overall accentedness in English and their pronunciation accuracy of the liquid consonants /ɹ/ and /l/ over time. They concluded that LOR effects were dependent on the participants' phase in the L2 learning process. That is, beginning learners showed better gains over time than more advanced learners over the same amount of time. Thus, as with AoA, LOR has been inconclusive as a definitive predictor of foreign accent.

L1 effects. Another interspeaker factor having been shown to influence L2 phonological acquisition is that of participant L1 and L1 use, and a considerable amount of research has revealed interlanguage effects on this process. Various models have been proposed suggesting that L2 phonological learning is systematic and follows stages whereby L1 and L2 cross-language similarities and differences affect progression through these stages. Major's (1987, 2002) ontogeny-phylogeny model of L2 phonological learning introduced three stages of L2 phonological learning: an initial stage heavily influenced by L1, a variable intermediate stage, and finally a stage represented by target-like L2 performance. Furthermore, Gatbonton;s (1975, 1978) gradual-diffusion framework argued that nontarget L2 forms often reflective of L1, depending on the context, are gradually replaced by target-like L2 forms. According to these models, L1 can have more of an impact on L2 phonological learning, at least in the beginning stages of learning.

In a study investigating 61 NNSs of English, Purcell and Suter (1980) found L1 to be the strongest predictor of pronunciation accuracy scores (R^2 = .42); they concluded that less L1 and more L2 usage correlated to more accurate L2 pronunciation. Similar investigations were done by Trofimovich, Gatbonton, and Segalowitz (2007) on 40 French-speakers learning English. The researchers examined how L1 might impact L2 pronunciation accuracy of the interdental fricative /δ/. Results indicated that accuracy depended on the initial, medial, final, or cluster contexts in which the phoneme was produced. More specifically, for French L1 speakers, English pronunciation accuracy was found to be more accurate in the sentence-initial and in the voiced fricative/affricate contexts and significantly different ($p < .001$) compared to the other three contexts.

Other studies investigating different L1s found L1 effects on accent and pronunciation accuracy in the L2. In their study of 60 Italian immigrants in Canada, Flege, Frieda, and Nozawa (1997) investigated how L1 usage affected English pronunciation accuracy. They found that the accent ratings for the English native-speaker controls were higher than those of the Italian NNSs of English whether they used a little or a lot of Italian in their daily lives. However, the participants who used more Italian had a significantly stronger foreign accent than those who used less Italian ($F(2, 57) = 10.49$, $p = .001$). Additionally, Flege et al. (1999) found that of their 240 Koreans participants, those who reported using English more in their daily lives had significantly better English pronunciation than those who reported using Korean more in their daily lives ($F(1, 38) = 4.27$, $p < .05$).

Gatbonton and Trofimovich (2008) discovered L1 and L2 effects in their investigation of 59 adult French-English bilinguals. They found that higher L2 proficiency was correlated to the amount of L2 use ($r = .37$ to $.53$, $p < .007$). Not only is target language use a potential factor impacting L2 phonological acquisition, but so is cultural identity. In Polat and Mahalingappa's (2010) study of 121 Kurdish middle and high school students living in Turkey, stronger foreign accent ratings were found for participants with stronger identification with the L1 community.

Conversely, the participants with greater identification with the L2 community resulted in more native-like L2 production.

Despite the above findings, other studies have reported little to no effect of L1 and L2 use on foreign accent. Thompson (1991) found that the amount of English use at home and the amount of English use with friends for the 36 Russian participants in her study only moderately correlated to better native-like accent ratings ($r = .33$, $p < .05$ and $r = .58$, $p < .01$ respectively). Similarly, working with 66 American learners of Spanish, Elliot (1995) investigated 12 variables, a couple of which implied increased L2 use, to ascertain whether they influenced pronunciation accuracy in Spanish. His findings showed that that the two factors of (a) having Spanish-speaking relatives and (b) traveling to Spanish-speaking countries had little effect on their pronunciation accuracy. In concluding, as with the other interspeaker factors of AoA and LOR, results of L1 and L2 use on foreign accent remain somewhat vague.

Ability to mimic. One of the models of L2 phonological learning proposed by Flege (1995) viewed the ability to differentiate segmental differences as a key component to this process. In his speech learning model, Flege argued that the ability to discern differences between L1 and L2 segments helps in the acquisition of L2 phonological accuracy. Another related interspeaker factor that has been investigated and warrants consideration in relation to the previous factors is that of the ability to mimic and its impact on pronunciation accuracy.

Early studies examining mimicry ability, such as Pike's (1959) study of 88 participants at a linguistic institute in the United States, created measurement tools to attempt to quantify participants' ability to imitate sounds. Both English and non-English segmental sounds and suprasegmental elements, such as pitch differences, were included in the mimicking instrument.

In Suter's (1976) study of 61 NNSs of English, an instrument adapted from Pike's instrument was created to evaluate the participants' ability to imitate English sounds across three attempts. An aptitude for oral mimicry (Suter, 1976) was assessed for each speaker based on their ability to imitate sounds on the first attempt while accounting for L1 effects by converting their scores based on the performance of others of the same L1. This aptitude for oral mimicry was later evaluated by Purcell and Suter (1980) in regression analyses and found to be one of the strongest predictors of pronunciation accuracy. In fact, for the 61 NNSs of English, aptitude for oral mimicry contributed to an additional 18% toward the variance in score, and combined with L1, the two held almost 56% of the variance in pronunciation accuracy scores, the best two-model predictor for foreign accent.

In Thompson's (1991) study of 36 Russian-born adult immigrants to the United States, the ability to successfully mimic sounds was also a factor influencing the pronunciation scores of the ESL participants. It added an additional 5% to the variance in scores with a combined R^2 of 82% along with AoA and gender. In their study of 240 Korean speakers of English, Flege et al. (1999) found the ability to imitate foreign accents as one of the principle factors (R = .783) influencing foreign accent ratings. In this study participants were asked to complete a self-evaluation questionnaire of their pronunciation abilities and then required to repeat sentences, which were later assigned a foreign accent score. Within the results from the questionnaire data, the ability to imitate foreign accents was grouped with the ability to remember how words are pronounced and musical ability to create a larger factor which they called "sound processing ability." This factor was the third strongest predictor of foreign accent, after the factors of AoA and LOR, and the three factors combined accounted for 74% of the variance in accent score.

Intraspeaker Factors

Other factors shown to impact the phonological acquisition process are intraspeaker factors, those that examine variability within language learners. This particular vein of research stems from sociolinguistics and dates back to the late 1960s where a number of studies began to examine intraspeaker variability as it relates to SLA and in particular phonological differences based on regional and social factors. Munro and Derwing (1994) outlined several intraspeaker factors that have been found to influence phonological accuracy: identity (Bailey, 2000; Beebe, 1977, Chambers & Trudgill, 1980; Labov, 1966), inhibition level (Guiora, Beit-Hallahmi, Brannon, Dull, & Scovel, 1972), emotional impacts of speech task (Dowd, 1985), and degree of self-monitoring (Dickerson, 1975; Dickerson & Dickerson, 1977; Labov, 1966).

Identity. NSs employ a variety of pronunciation choices based on identity factors. Labov (1966) noticed the different speech styles employed by New Yorkers as a result of various social class settings. More specifically, he observed the production of the post-vocalic /ɹ/ across three different social classes and within the participants. Variability within the participants, in particular, showed that the environment in which the speech occurred played a role in whether or not to pronounce the post-vocalic /ɹ/. Chambers and Trudgill (1980) similarly investigated the phonological feature of post-vocalic /ɹ/ production in another environment, Great Britain, and also found intraspeaker variability based on identity factors like social class and prestigious interpretations. These two often cited studies found contradictory results based on their settings; that is, in New York, the presence of post-vocalic /ɹ/ increased as social class increased while in England, the presence of post-vocalic /ɹ/ decreased as social class increased.

NNSs also employ different dialectal and phonological choices based on their environments. Sociolinguistic inquiry into identity factors influencing SLA have examined identity under the poststructuralist framework (Pavlenko & Blackledge, 2004; Norton, 2013). Under this theoretical framework, individuals are seen as negotiating multiple identities (e.g. gender, ethnic, social class, professional) in their daily interactions. Language choices in grammatical, lexical, and phonological features serve as just some of the strategies used to negotiate these identities. For example, Bailey's (2000) study of a Dominican teenager showed how this language learner code-switched from Spanish English to Black English to emphasize his social identification with both his Dominican identity and his blackness.

Other studies have investigated identity factors of immigrants living in another country to see how these identity roles influence SLA (Block, 2006; Goldstein, 1996; Norton Pierce, 1995; Norton, 2000; Teutsch-Dwyer, 2001). These studies have shown how L2 use and accuracy, including pronunciation, are influenced by the particular identity being negotiating "across time and space" (Norton Pierce, 1995, p. 18) In Beebe's (1977) study of 17 Chinese-Thai bilingual children, phonological variation within the participants depended on the identity of the interlocutor; more specifically, when they were speaking with a Thai listener, as compared to speaking with a Chinese listener, the nine phonological features under investigation were produced with the Thai variant for those sounds. Thus, the findings showed that learners adapted their phonological productions based on their social identities interacting with those of whom they were speaking. Gatbonton and Trofimovich (2008) studied ethnic group affiliation as the identity factor influencing L2 proficiency. For the 59 French-English bilinguals, those with stronger francophone EGA scores correlated with stronger perceived foreign accent scores in

English. The researchers concluded that identify factors, such as ethnic group affiliation, caused variability both across and within language learners.

Affective factors. Besides identity, inquiry into other intraspeaker effects on L2 phonological accuracy include the degree of inhibition of the interlocutor and the emotional impact of the topic. Looking at L2 phonological acquisition, Guiora, Beit-Hallahmi, Brannon, Dull, and Scovel (1972) situated pronunciation as the most important aspect of language ego in that it tends to be the hardest to acquire in an L2. The 87 native English speaking participants in their study who ingested small amounts of alcohol had an increased accuracy in their L2 pronunciation in Thai. Thus, this study demonstrated facilitation in phonological accuracy when inhibition levels were lowered. It also showed how language ego can be a barrier to L2 phonological acquisition. Schumann (1975) suggested that other affective factors, such as ego permeability, acculturation, attitude, and motivation, might be more important than maturational factors for adults learning another language. In his acculturation model (Schumann, 1986), he argued that the acculturation process in the target culture played a vital role in acquiring the new language.

Another study investigating intraspeaker variability and phonological accuracy was Dowd's (1985) study of 100 adult Mexican learners of English. Her focus was on whether a shift from an unemotional task to an emotional task within a single interview would produce a decrease in target language accuracy of six phonological variables. Her findings showed that, with the emotional task, only two of the six phonological features decreased in phonemic accuracy, while the other four phonological features actually increased in accuracy. Thus, Dowd concluded that this shift in emotional task did not necessarily result in a shift toward more or less interlanguage vernacular forms.

Other types of tasks with varying emotional effects were also examined in a bidialectal setting. Wassink, Wright, and Franklin (2007), for instance, investigated 10 bidialectal Jamaican mothers on the variability in vowel productions of their Jamaican Creole and Jamaican English depending on four different speech environments: speech directed at infants, speech delivered under background noise, speech produced with the intention to ease observed listener difficulty, and speech produced in a more formal experimental environment. The results showed intraspeaker variability in English tense-lax vowel distinctions depending on the elicitation procedure. The authors concluded that their participants' adjustments resembled those of other sociolinguistic studies of style-shifting (Chambers & Trudgill, 1980; Labov, 1966).

Monitor hypothesis. Besides affective factors potentially influencing phonological acquisition within a single language learner, the complexity of the task and subsequent degree of self-monitoring may also bear some responsibility. In a related study, Borden, Gerber, and Milsark (1983) examined 10 adult Koreans learning English and found intraspeaker differences based on the complexity of the speech task; that is, the pronunciation of both /ɹ/ and /l/ was judged to be worse in tasks involving semantic, syntactic, and lexical components as compared to those involving speaking nonsense syllables. Thus, the authors concluded that more complex speech tasks negatively impacted L2 phonemic accuracy, so when speech is elicited for the purposes of evaluating pronunciation accuracy, task complexity needs to be considered.

Krashen's (1978) Monitor Model Hypothesis suggests that L2 learners sometimes monitor, or utilize more conscious knowledge, during their L2 speech production. This monitoring, according to Krashen, occurs with variation across different speakers and within the speakers themselves. Krashen (1982) argued that some L2 speakers are more “extreme Monitor users” while performing, while others hardly ever monitor their output (p. 12). Krashen outlined three characteristics of Monitor users: (a) Monitor users are able to edit their L2 output successfully only when done without interfering with communication, (b) monitoring results in variation in performance depending on different tasks and conditions, (c) monitor users are concerned with accurate language production and view unmonitored production as “careless” (p. 12-13).

Several SLA studies have sought to understand which tasks and situations might lead to more optimal monitoring and have found an increase in phonological accuracy with tasks involving an increase in self-monitoring (Beebe, 1980; Dickerson, 1975; Dickerson and Dickerson, 1977). Studies have juxtaposed pronunciation accuracy following read-aloud tasks versus free response tasks and have shown that more formal contexts, such as reading, involve increased monitoring and higher accuracy. For instance, Dickerson (1975) examined variability of /z/ in 10 Japanese speakers learning English at an American university over a nine-month period. She analyzed their speech inside three speech elicitation tasks: a free-speaking, a reading of dialogues, and a reading of a word list. Her participants showed greater accuracy of /z/ in the more formal tasks of reading word lists and reading dialogues as compared to the more informal task of free speaking.

Testing the same three elicitation methods but this time on /ɹ/, Dickerson and Dickerson (1977) again reported fewer errors in the production of the English /ɹ/ for the Japanese learners of English when reading word lists or dialogues as compared to when spoken spontaneously.

Similarly, Beebe (1980) examined the production of /ɹ/ in initial and final singletons and clusters for nine adult Thai participants. Speech was elicited via four different methods: an interview conversation, a passage reading, a word list reading taken from the passage, and a listening perception test. Beebe reported that the final /ɹ/ was more accurate in the more formal reading tasks due to what she claimed involved more optimal monitoring conditions than the informal conversation tasks; however, the initial /ɹ/ was more accurate in the informal task.

Liu and Fu (2011) and Liu (2011) examined 10 problematic sounds for 60 Chinese speaker students majoring in English at a Chinese university. The participants were asked to complete three speaking tasks of vocabulary reading, sentence reading, and simultaneous speech in a pre and posttest design. Findings suggested negative transfer effects in both pre and posttests, as well as an impact of task type on phonemic accuracy. Attention to accuracy, while often improving segmental accuracy, was found to impact fluency. Liu (2011) attributed this finding to most of the participants' being monitor over-users, exaggerating the learned sounds in unnatural fluency in their productions.

Other studies have found different results as to whether reading tasks involve more optimal monitoring and, thus, greater accuracy in L2 production. Oyama (1976) found less accuracy for the 60 Italian participants in the reading tasks and claimed that reading tasks resulted in more emotional stress as compared to recounting a story, which tended to be treated more informally. Oyama suggested that reading resulted in more accented speech because of the demands of the task itself. Thompson (1991) also found similar findings for the 36 Russian participants of more accented speech in the speech samples elicited from reading passages compared to the spontaneous speech samples in her study. However, in Munro and Derwing's

(1994) study of 10 Mandarin speakers learning English, participants were asked to give a narrative account of a series of cartoon, and later they were asked to read a transcribed portion of their own narrative story. The researchers did not find statistically significant differences in foreign accent scores based on speaking or reading tasks.

With regard to the factors influencing L2 phonological acquisition, the studies have explored both external and internal factors that appear to interact with the goal of some language learners of an ultimate level of phonological attainment. Whether or not there is a critical or sensitive period for phonology is not certain as there appears to be great variety between individual language learners, with some learners able to obtain native-like pronunciation and others not. Interspeaker factors, including maturational and sociocultural characteristics, seem to influence pronunciation accuracy. Moreover, factors within language learners themselves influence their accuracy on any given day, situation, or context in which the language is used.

Functional Load

When phonemic accuracy is under investigation, it is important to consider the content of what participants are being asked to speak. Functional load refers to "the extent and degree of contrast between linguistic units, usually phonemes" (King, 1967). Research in child L1 acquisition has led to discoveries regarding the functional load and its impact on the order of phoneme acquisition (Pye, Ingram, & List, 1987; Van Severen, Gillis, Molemans, Van den Berg, De Maeyer, & Gillis, 2013). While it is not the only factor at play, phonemes involved in high functional load oppositions tend to be acquired before others (Van Severen et al., 2013). In SLA research, both Brown (1988) and Catford (1987) attempted to systematically describe the most frequently occurring phonemes and phonemic contrasts in English. Catford identified the most frequently occurring phonemes as the number of times the phoneme or a phoneme contrast occurred per thousand words of text. He suggested a hierarchy in pronunciation instruction guided by the more frequently occurring phonemes, more specifically, that two principles should guide the selection of sounds for pronunciation instruction – *frequency of occurrence* and *functional load*.

Language frequency became easier to quantify once the field of corpus linguistics developed. Reppen, Fitzmaurice, and Biber (2002) stated that "adequate descriptions of variation and use must be based on empirical analysis of natural texts" (p. vii) and not just on linguists' intuitions of language variation and use. Corpora might be better capable of handling the quantity of language needed to truly understand language in use in society as they offer objective measures of vocabulary and grammar frequencies (Biber & Reppen, 2002). Catford (1987) further argued that "the *principle of frequency* often used in L2 vocabulary selection can also be applied to pronunciation" (p. 88). Thus, the frequency of phonemes based on corpora of spoken English could serve to guide the research of pronunciation.

Just as the Brown Corpus (Kučera & Francis, 1967) was being developed, others started investigating in a more systematic way the notion of frequency of occurrence of English sounds. A study looking at frequencies of spoken English sounds found that the most frequently occurring vowel sound is the schwa /ə/, followed by the front high vowel /ɪ/ (Denes, 1963). In fact, in Denes' study, the three most frequently occurring vowel sounds – /ə/, /ɪ/, and /aj/ – accounted for half of all vowel occurrences (p. 894) and 75% of all productions in stressed syllables (p. 898). Furthermore, taking the top two vowels of /ə/ and /ɪ/ into consideration, their frequencies were higher in unstressed syllables (p. 897). As for consonants, Denes reported the most frequently occurring consonant sounds, accounting for half of all consonant occurrences, were the alveolar sounds /t/, /n/, /d/, /z/, /s/, and /l/, with the most frequently occurring consonant in initial position being the /δ/. These findings helped inform actual phonemic distribution in spoken, rather than written, English.

Looking solely at frequency of occurrence, Catford (1987) claimed the most frequently occurring English consonant sound as the /δ/ (p. 85). Looking at phonemic contrasts, he reported that higher word frequencies exist in the English language for /i/-/ɪ/ as compared to /u/-/ʊ/ minimal pairs. In other words, the /i/-/ɪ/ phonemic contrast, seen in words like *peak* and *pick*, serves to differentiate the most word pairs, and as such possesses a high relative functional load. On the other hand, much fewer words distinguish the /u/-/ʊ/ contrast, seen in words like *fool* and *full*, and thus the /u/-/ʊ/ phonemic opposition is said to possess a low relative functional load. Catford argued that the selection of sounds for pronunciation teaching should be informed by their frequency of occurrence and relative functional load with sounds and phonemic contrasts possessing higher FLs taking priority in pronunciation instruction over lower FLs (Catford, 1987).

Munro and Derwing (2006) tested the functional load hypothesis on recordings of 80 Cantonese-accented speakers of English to see whether high functional load errors held a greater impact on listener judgements of accentedness and comprehensibility as compared to low functional load errors. Utilizing accentedness scales from 1 (no accent) to 9 (strong accent) and comprehensibility scales from 1 (very easy to understand) to 9 (very hard to understand), they concluded that high functional load errors resulted in a greater degree of perceived accentedness than low functional load errors; in addition, they found a significant effect of high versus low functional load error condition (*F*(6, 72) = 31.09, *p* < .001) on comprehensibility ratings. This study was one of the first to support the functional load hypothesis as a potential framework for prioritizing segmental errors and for focusing pronunciation research and teaching efforts.

Kang and Moran (2014) examined the functional load approach in how phonological errors affect oral assessments. In particular, they looked at 120 speech samples from the Cambridge ESOL General English Examinations, an English language qualification exam, to see which segmental features deviated from SAE. Findings showed statistically significant differences across proficiency levels for high relative functional load consonant errors ($F(3,111) = 8.65$, $p = .000$) and high relative functional load vowel errors ($F(3,111) = 12.36$, $p = .000$) but not for low relative functional load errors. Thus, the authors stressed the importance of considering the functional load as a factor to include in prioritizing pronunciation features to isolate in research studies.

Speech Elicitation Methods

In addition to evidence supporting which phonological features to include in pronunciation research, how to elicit those features is equally important to consider. When determining if there is a more effective way or order to elicit speech, the field of cognitive psychology has done considerable research on the mind and how different perceptual modalities, such as auditory, visual-orthographic, and visual-non-orthographic, are processed and how they interact with each other. One well-known example of how visual to orthographic stimuli compete with each other can be seen in the Stroop task (Stroop, 1935). According to Stroop (1935), the brain has difficulty processing orthographic representation of color words when those words are color-coded in different colors than their respective names.

Stuart and Carrasco (1993) conducted three experiments with Stroop-like tests using either pictures or auditory words as distractors for their naming tasks. They found that word distractors stemming from the same semantic category interfered more with naming the pictures than word distractors from semantically unrelated categories; thus, they reported a semantic component to the Stroop-like effect. The Stroop effect was later examined with highly proficient Spanish-Catalan bilinguals to determine if there was a lexical competition across the two languages. Two effects, the cross-language identity effect and the phono-translation effect, were tested via Stroop tasks. Results showed the presence of the cross-language identity effect but no phono-translation interference effect from the Stroop tasks, thus weakening the evidence for lexical competition across the languages (Costa, Albareda, & Santesteban, 2008).

Brain research conducted in the field of adult psycholinguistics on lexical processing has almost universally shown that L1 lexical concepts are recognized before their corresponding phonemes and phonological forms are identified (Indefrey & Levelt, 2004; Levelt, Roelofs, & Meyer, 1999; Schmitt, Münte, & Kutas, 2000). The underlying processes of speech production are complex, but Levelt, Roelofs, and Meyer (1999) theorized four main lexical processing stages in order of processing time: (a) conceptual preparation, (b) lemma retrieval, (c) form encoding, and (d) articulation of the word. Their computational model was constructed after a series of reaction time experiments, primarily of picture naming and related word productions.

Further testing this model, Schmitt, Münte, and Kutas (2000) examined L1 processing time of phonological encoding during picture elicitation in German. Of the 120 black-on-white drawings, half began with a consonant and half with a vowel, and participants had to indicate the starting sound of the picture viewed as either a consonant or a vowel. Results again showed faster response times for semantic processing over phonological processing. Furthermore, Indefrey and Levelt (2004) conducted a meta-analysis of 82 word-production and auditory-perceptual studies and found shorter processing times for conceptual processing and lexical access than phonological code retrieval. In other words, concepts are recognized more quickly than they can be pronounced. These findings of L1 brain research may demonstrate in L2 studies a toggling of input across orthographic and non-orthographic modalities.

Word Reading Tasks

Morton (1969) outlined a model for word recognition, the basic unit of which he called a *logogen*, or word generation. In this model, each logogen was defined by three sets of attributes: semantic sets, visual sets, and acoustic sets. According to Morton, word information received during reading and listening is continually being processed in this model before output can occur. Morton's (1980) logogen model underwent various embellishments when he added grapheme-phoneme and acoustic-phonemic routes to output, thus supporting the notion that written and spoken words operate under distinct recognition systems. Morton argued that some important properties of word identification through reading tasks include regularity of letter-sound correspondences and frequency of usage (Morton, 1969).

Furthering the notion of cognition processes in reading, Coltheart, Curtis, Atkins, and Haller (1993) investigated the dual-route cascaded (DRC) model of word reading, which involved two models for converting printed text to speech that are either lexical or letter-to-sound based. Coltheart, Rastle, Perry, Langdon, and Ziegler (2001) later promoted the DRC model as the only computation model that can perform both lexical decision and reading aloud tasks. In alphabetic languages, like English, letter-sound regularity has shown to influence processing of low-frequency words but not of high-frequency words; in other words, the pronunciation of the more frequent words was retrieved more quickly from the lexical store than via grapheme-mapping.

In Stroop-like tasks, words have been found to be faster than pictures at accessing articulatory information (Smith & Magee, 1980). Looking further into L1 acquisition, Rosenthal and Ehri (2008) found positive effects of orthography in two separate experiments with 20 minority second graders and 32 minority fifth graders in the United States. Participants were shown the written forms of one set of words and not of another set of words. Results showed that having the orthographic form enhanced memory of word meanings and pronunciation of words. More specifically, participants were able to recall the spoken forms of new words better if the words had been presented the day before with spellings. Since spellings were presented alongside with pronunciations, they interpreted the results as being due to a different form of decoding, that of creating an orthographic map which connected grapheme and phoneme representations of the words. Thus, while orthography had been shown to aid with vocabulary learning, in this study it also helped with pronunciation accuracy.

Cross-linguistic studies have categorized different languages as possessing varying degrees of orthographic depth whereby languages with high letter-sound correspondence are said to have shallow orthographic depth, while those with less regular correspondence are said to have deep orthographic depth. Jiang (2016) contrasted four L1s of differing orthographic depth, Spanish, Chinese, Arabic, and Japanese, to see whether oral reading fluency predicted ESL reading comprehension based on the different L1 backgrounds of the 149 ESL participants. Findings revealed that prosody was the only significant predictor for Chinese (R^2 = .205) and

Japanese (R^2 = .339) participants, oral reading efficiency was the only significant predictor for Arabic participants (R^2 = .471), and word accuracy and oral reading rate were both significant predictors for Spanish participants (R^2 = .555 and R^2 = .671 respectively).

In looking at how L1 orthographic features influence L2 word recognition processing, Akamatsu (1999) examined 50 highly-proficient ESL readers, with Test of English as a Foreign Language (TOEFL) internet-based reading section scores between 57-68, from China, Japan, and Iran to see whether word case alternation (i.e. wOrD cAsE) affected reaction times for a naming task. Both high-frequency and low-frequency words were selected and presented in two sets, one in lower case and the other in alternated case. The author reported a statistically significant difference in response times based on participant L1 ($F(3, 63) = 34.92$, $p < .0001$) as compared to the NS controls, with Persian speaking participants, whose L1 is alphabetic, as less affected by case alternation than Chinese and Japanese speaking participants, whose L1 is nonalphabetic ($F(1,63) = 4.23$, $p < .05$). Furthermore, when accounting for word frequencies and response phonological accuracy, the Chinese and Persian participants revealed a significantly higher error percentage in high-frequency words compared to low-frequency words ($F(3,456) = 8.82$, $p < .0001$). The author attributed this finding to potentially different word-recognition mechanisms at play for alphabetic and nonalphabetic words.

Wang and Koda (2007) also tested the naming accuracies of English words with 18 Chinese and 16 Korean ESL participants. These words were grouped into high-frequency and low-frequency words with regular and irregular spellings and nonwords. Participants were shown all words on a computer screen and asked to read them out loud. They found an L1 group effect with Korean adults having overall better naming accuracies than Chinese adults of both real words and nonwords ($F(1, 32) = 21.67$, $p < .001$). Wang and Koda attributed their findings to better performances in English reading of words out loud coming from participants with more alphabetic L1s (Korean) compared to nonalphabetic L1s (Chinese).

Other SLA studies have revealed different findings as to the influence of orthography on pronunciation accuracy, with some showing beneficial effects and others no effects. In three separate studies, Simon, Chambless, and Alves (2010) trained 20 NSs of American English with no formal instruction in French or German to pronounce a vowel contrast absent in their native language. The adult participants were trained in one of two conditions, auditory only and auditory linked with spelling, and findings revealed that the participants trained with orthographic forms did not outperform participants trained without orthographic form; thus, there was a lack of orthographic effect on learning a nonnative vowel contrast.

However, Erdener and Burnham (2005) found facilitative effects of orthography on nonnative speech production. Focusing on visual speech with and without orthographic information, 32 NSs of Turkish (shallow orthography) and 32 NSs of Australian English (deep orthography) were tested with nonwords in Spanish (shallow) and Irish (deep) under four conditions: auditory only, auditory visual, auditory-orthographic, and auditory-visual-orthographic. The visual representation of words was done via a videotaping of a NS's face while speaking the words. Results derived from an analysis of phoneme errors revealed a facilitative effect of having orthographic information present ($F(1, 61) = 36.788$, $p < .01$) for all groups regardless of the L1, hence the auditory-orthographic and the auditory-visual-orthographic conditions.

Word Repetition Tasks

The ability of humans to repeat what they hear has been investigated for some time in the field of psycholinguistics. Baddeley and Hitch (1974) suggested two storage sectors, or loops, for storing information temporarily in working memory: the visuospatial sketchpad and the articulatory, or phonological, loop. They argued that the visuospatial sketchpad receives and processes visual input while the phonological loop does the same for auditory input. Baddeley (1986) later developed the idea of the phonological loop, whereby he maintained that the retention of verbal information comprised a phonological store and a rehearsal process. According to Hamada and Koda (2011), "adult L2 learners are likely to learn new words both orally and visually at the same time" (p. 76), so access to the phonological loop plays a key role in English word learning. Because the ability to mimic sounds has been found to be one of the predictors of pronunciation accuracy in SLA studies (Flege et al., 1999; Purcell & Suter, 1980; Thompson, 1991), it is important to consider word repetition as a valid elicitation method in pronunciation research and assessment and to compare it to other elicitation methods.

One study that found a facilitative effect of word repetition on L2 pronunciation accuracy was Elliott's (1997) study of 66 undergraduate American learners of Spanish. Participants were asked to repeat words and sentences, read words, and perform a picture description task via pre and posttest instruments focused on several difficult segments. Results indicated improvements in the segmental productions in the word repetition task ($F = 11.16$, $p < .001$), the sentence repetition task ($F = 15.11$, $p < .0001$), and the word reading task ($F = 17.07$, $p < .0001$), but not in the free elicitation task.

Research also exists in word repetition that involves multiple utterances of the same word, also known as word priming. The benefits of word priming have been found in native speaker studies (Buchwald, Winters, & Pisoni, 2009; Kim, Davis, & Krins, 2004). Examining word priming of foreign words, Service, Yli-Kaitala, Maury, and Kim (2014) conducted a study of 28 Finnish adults and 27 Finnish children on their repetition of Korean words. Participants were asked to repeat 82 polysyllabic words selected from Korean textbooks, some presented once and others presented five times. Although there was no statistically significant difference between the adult and children groups in terms of pronunciation accuracy, greater accuracy in pronunciation was found within both study groups for the words repeated five times compared to those only repeated once ($F(1, 47) = 8.335$, $p < .01$).

Outside of SLA, the field of speech and language pathology has also investigated the effects of auditory elicitation. Investigating 12 Spanish-speaking children with phonological disorders, Goldstein, Fabiano, and Iglesias (2004) examined the relationship between spontaneous and imitated speech for phonological disorder analysis. Results indicated that consonant accuracy varied as a function of the elicitation method; more specifically, while 62% of the words were produced identically between the two tasks, the imitated task resulted in more accuracy in 25% of the cases, suggesting that imitated responses should be incorporated as an elicitation method for phonological sampling.

Picture Naming Tasks

When picture naming has been investigated via Stroop-like tasks in L1 studies, results have shown that pictures are faster than words at accessing semantic information (Smith & Magee, 1980). Eliciting speech from picture naming tasks has also been studied in L1 acquisition. While the focus of the study was related to native-speaker semantics and not pronunciation accuracy, Snodgrass and Vanderwart (1980) presented a set of 260 images to elicit speech. The black and white illustrations were presented to 219 native English speakers who then wrote the first name that came to their minds. During the data analysis process, the names given by the participants were ranked, and, each word was assigned a name agreement percentage. The Snodgrass and Vanderwart list underwent various revisions and additions over the subsequent 30 years and with a variety of other languages: English (Adlington, Laws, & Gale, 2009; Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010), Spanish (Moreno-Martinez & Montoro, 2012), and Italian (Viggiano, Vannucci, & Righi, 2004). In particular, the work done by Moreno-Martinez and Montoro (2012) not only changed the 260 Snodgrass and Vanderwart illustrations to color, but also added another 100 color images. The researchers also analyzed data for even more psycholinguistic variables than Snodgrass and Vanderbilt had investigated, such as age of acquisition. Succinctly put, picture naming studies have shown that more familiar items, those with higher name agreement, tend to be easier to name.

Another native-speaker study (Qu, Zhang, & Damian, 2016) of literate Chinese-speaking adults investigated the differences between picture naming and written word naming on lexical access time measured in milliseconds. The results of this study showed that both written and spoken word production shared similar lexical access times. The factor that most affected time was that of word frequency, high versus low-frequency words. The authors concluded that both written and spoken object naming, therefore, followed similar levels of cognitive processing.

In further examination of language phenomena, the field of speech and language pathology has investigated picture naming tasks and their effectiveness in assessing speech-sound disorders. To get a more accurate picture of a person's phonological attainment and behavior, two elicitation methods have been compared: single word naming from picture naming tasks and extemporaneous speech from conversation tasks. In one such study comparing the two methods of picture naming and conversational speech, statistically significant differences were found for the 61 speech-delayed children participants (Morrison & Shriberg, 1992). More specifically, established sounds were produced more accurately in the conversation task, while emerging sounds were produced more accurately in the picture naming task.

In a study by Wolk and Meisler (1998), the picture naming task yielded more phonological errors as it tapped more deeply into the phonological system of 13 phonologically impaired children in the study. This study confirmed several strengths of picture naming over spontaneous conversation as elicitation methods for assessing phonological behaviors: predetermined word lists presented via pictures are simpler to use, easier to control, and more facilitative to analyze over time. However, in order to have a more balanced understanding of a person's phonological abilities, the authors suggested that both types of elicitation methods, picture naming and conversation, be used whenever possible.

It is not surprising that studies investigating picture naming abilities of bilinguals have shown slower naming times in the weaker language compared to the stronger language (Ivanova & Costa, 2008; Kroll, Bobb, & Wodniecka, 2006) and that the bilingual naming delay continues despite increased proficiency in the other language (Sholl, Sankaranarayanan, & Kroll, 1995). Hanulová, Davidson, and Indefrey (2011) attempted to understand where in the lexical processing stage the delay occurred, and they suggested frequency and interference effects occurring in one of the postlexical stages of morpho-phonological code retrieval, phonological encoding syllabification, or phonetic encoding. Lastly, in a study investigating cross-language picture naming, Spanish-English and Japanese-English bilinguals picture naming speed was facilitated by cognates in English (Hoshino & Kroll, 2008). The authors concluded that "there is cross-language activation of phonology even for different-script bilinguals" (p. 501).

Measurements of Phonemic Accuracy

In SLA, traditional approaches to pronunciation research were mostly phoneme-based and restricted to segmental practice and word-level accuracy. Different approaches to teaching English, namely audiolingualism, brought pronunciation to the forefront of research efforts while other approaches, such as the communicative language teaching approach, have primarily diminished the importance of explicit pronunciation teaching (Celce-Murcia, Brinton, & Goodwin, 1996). How phonemic accuracy has been measured varies depending on the focus used to investigate L2 pronunciation.

Principles of Intelligibility and Nativeness

Two very different principles have guided L2 pronunciation research, the *intelligibility* principle and the *nativeness* principle, each with roots stemming from different language learning approaches. Levis (2005) offered definitions of these contradictory principles. The intelligibility principle posits that language learners merely need to obtain speech that is comprehensible in the target language, while the nativeness principle claims that with effort and motivation language learners might be able to obtain native-like pronunciation in the target language (p. 370).

The focus on meaningful language, language that is intelligible, was espoused by language teaching approaches and SLA research in the early 1980s. The idea that L2 learning involves receiving comprehensible input (Krashen, 1982) mirrors that of L1 acquisition. Moreover, the grammar-translation method and other non-communicative approaches were superseded by a move from focusing on accuracy. This shift from focusing on accuracy to focusing on meaning brought a new face to pronunciation teaching and research. Linguists and educators started to question whether it was necessary to teach native-like accuracy in pronunciation, particularly when what constitutes “accurate” may be open to debate considering the growth of literature regarding English as an international language.

Since English has increasingly become a lingua franca and is viewed as an international language (Crystal, 1997), a variety of regional and social differences in English are abundant. Jenkins (2002) argues that accent differences should be endorsed in pronunciation instruction with more focus on intelligibility and comprehensibility rather than on pronunciation accuracy. Derwing and Munro (2009) add that if accent studies do not focus on intelligibility and comprehensibility, then instruction may be lost on areas not needed by language learners. While intelligibility principles are more endorsed nowadays (Derwing and Munro, 2009), both intelligibility and nativeness principles are still valued in SLA research.

Despite the trend in pronunciation research to focus on intelligibility, many language learners still aspire to learn the target language with native-like perfection, (Chiba, Matsuura, & Yamamoto, 1995; Dalton-Puffer, Kaltenboeck, & Smit, 1997; Scales, Wennerstrom, Richard, & Wu, 2006; Timmis, 2002). Unlike the intelligibility principle, the nativeness principle not only appreciates the attention-to-form embraced by the grammar-translation method but also the attention-to-accuracy championed by the audiolingual method of language learning. Under the nativeness principle, learning from lists of decontextualized words, which fell out of favor in communicative learning classrooms, is encouraged.

It has been argued that suprasegmental features impact comprehensibility more than segmental ones (Kang, Rubin, & Pickering, 2010), yet both segmental and suprasegmental pronunciation features influence comprehensibility, intelligibility, and accuracy. In fact, segmental errors can severely impact comprehension (Fayer & Krasinki, 1987), but which segmental errors differentiate learners' proficiency is not known. Indeed, if language learners hope to improve their pronunciation of the target language – whether for accuracy or intelligibility – phonemic accuracy is one key component to that goal.

Phonemic Accuracy Judgment

L2 research studies, guided by both intelligibility and nativeness principles, have utilized various measurement tools to interpret nonnative speech, including self-analysis, computer analysis, and rater judgments to measure accentedness, intelligibility, and phonemic accuracy. Because Derwing and Munro (2009) operationalized accent in terms of listener perception, it is not surprising that the majority of phonemic accuracy judgments have been done through native-speaker rater judgments. Rater judgment studies examining degree of foreign accent typically utilized accentedness or intelligibility scales along the spectrum of "native speaker" to "definitely foreign" or "very intelligible" to "not at all intelligible." Most of these studies have created 5-point scales (Bongaerts et al., 1997; Elliott, 1995; Oyama, 1976; Piper & Cansin, 1988; Thompson, 1991) while others have used 4-point scales (Asher & Garcia, 1969, Flege et al., 1997), 6-point scales (Moyer, 1999), and even 9-point scales (Flege et al., 1999, Gatbonton & Trofimovich, 2008).

Other rater judgment studies have focused more on specific phonemes and their L2 productive accuracy. These studies typically investigated one L1 and certain problematic phonemes in the L2. Phonological "errors" in these studies were often coded as deviations from native speaking norms and not as inaccurate or incorrect pronunciations. For instance, a number of studies have investigated the problematic phonemes of /ɹ/ and /l/ with Japanese speakers. Bradlow, Pisoni, Akahane-Yamada, and Tohkura (1997) conducted four experiments with 11 Japanese adult participants that looked at both perception and production of the two phonemes. Ten American English listeners evaluated the production of two tokens (pretest and posttest) of each phoneme on a 7-point scale indicting which sounded better. While the participants made improvements from the pretest to posttest minimal pair productions ($t(8) = -7.392$, $p < .005$), the researchers also found high variability in the results suggesting that the two processes of perception and production appear to operate under distinct domains, and improvement in one does not necessarily indicate improvements in the other.

Riney and Flege (1998) also examined the accuracy of /ɹ/ and /l/ for Japanese learners of English. In three different experiments, Riney and Flege had 11 adult native Japanese speakers and five native English speakers read sentences with 84 singleton and cluster target sounds. Although no specific rater statistics were presented, each experiment handled rater judgments in different ways. In the first experiment, five NSs judged the sentences for degree of foreign accent on a nine-point scale. In the second experiment, three trained NSs were presented with the first part of a word and were asked to identify which phoneme was uttered by clicking on one of three boxes labeled /ɹ/, /l/, or neither. In the third experiment, 10 untrained NSs were presented with pre and posttest demisyllables of both /ɹ/ and /l/ and asked to indicate which of the two sounds contained the better exemplar of the intended sound by clicking on one of two boxes labeled on a computer.

Similar dichotomous ratings were utilized in the Liu and Fu (2011) and Liu (2011) studies of 60 native-speaking Chinese English majors. The 10 target sounds were given a score of either zero for inaccurate or one for accurate on the 10 target sounds by a highly qualified teacher. Trofimovich, Gatbonton, and Segalowitz (2007) also examined the pronunciation accuracy of the voiced interdental fricative /δ/ with 40 Francophone English learners in Canada. Speech was gathered from a 440-word reading passage that included 80 target tokens of /δ/ in eight phonetic contexts. Ten native English listeners, all with English teaching experience, analyzed the pronunciation accuracy of each speech sample via a binary decision of an accurate or inaccurate production of /δ/. Interrater reliability statistics were calculated in this study with moderate to high levels of reliability (α range of .70-.99).

Two Canadian studies investigated both global English accent ratings and L2 pronunciation accuracy of 48 NNS adults enrolled in ESL programs. In both studies, nonnative speech was elicited in two tasks, a narration of the cartoon story and a sentence reading task of high-frequency lexical items. Derwing, Munro, and Wiebe (1998) utilized 48 NSs who rated the speech samples for comprehensibility and accent on nine-point scales. Pearson coefficients (r) of .71 on comprehensibility and .70 on accent were reported for interrater reliability. Derwing and Rossiter (2003) replicated the earlier study design but this time with six NS judges, all with extensive ESL experience, who rated the speech samples on nine-point scales for accentedness and comprehensibility; here, interrater reliability indicated similar coefficients of .69 on accentedness ratings and .72 on comprehensibility ratings.

Despite the attempt of many researchers to establish intra- and interrater reliability in their rater judgment studies, some factors have been shown to weaken the effectiveness of these evaluations of nonnative speech. Rater bias has been shown in: rater attitudes toward specific L1 groups (Lippi-Green, 1997; Yook & Lindemann, 2013), rater familiarity with accents (Chiba, Matsuura, & Yamamoto, 1995; Dalton-Puffer, Kaltenboeck, & Smit, 1997; Rajadurai, 2007), rater exposure to languages (Baetens-Beardsmore, 1979), rater preference based on speaker traits of solidarity and status (Bayard & Green, 1995; Chong & Tan, 2013; Giles, 1970), and rater preferences for certain identity categories (Bresnahan, Ohashi, Nebashi, Liu, & Shearman, 2002).

Another factor to consider is that of rater bias based toward accent variety and dialectal differences. Goldstein and Iglesias (2001) examined 54 Spanish speaking children of General Spanish and Puerto Rican Spanish to determine if there was an effect of dialect on the test results and found that there were some phonological patterns, such as final consonant deletion, for instance, affected by dialect ($F(1, 103) = 2030.65$, $p < .000$). An example provided was in the pronunciation of *dos* (*two*), the last consonant of which is often not pronounced in the Puerto Rican dialect. Thus, there were rater biases involved as the number of inaccuracies increased when dialect was not considered.

According to Schmidt and Sullivan (2003), the two most frequently used pre and posttests employed in accent modification courses in the United States are the Compton's Phonological Assessment of Foreign Accent (CPAFA; Compton, 1983) and the Proficiency in Oral English Communication Screening (POEC-S; Sikorski, 1991). These two assessments elicit speech of L2 learners of English both in isolated words and in extended speech, which is then transcribed at the phonemic level, typically by a speech and language pathologist or TESOL specialist. Morton, Brundage, and Hancock (2010) ran reliability and validity measures on the POEC-S test by comparing it to TOEFL scores and accent ratings of skilled and unskilled listeners. They found that skilled listeners had overall higher intrarater reliability with coefficients ranging from .82 on accent scores to .95 on intelligibility scores; the unskilled listeners also held moderate to high intra-rater reliability with coefficients ranging from .74 to .95 for all parameters assessed. In addition, the POEC-S total scores held strong correlations to TOEFL scores ($r = .78$, $p < .01$). Thus, the authors found the POEC-S test to have construct, criterion, and social validity for the 28 nonnative English speakers they tested.

Aside from assessing a foreign accent, the field of speech and language pathology also actively engages in measuring speech impairment via a number of published articulation and phonological assessments that are utilized to find error patterns and disorders within the English sound system. These include the Clinical Assessment of Articulation and Phonology (CAAP-2; Secord & Donahue, 2014), the Diagnostic Evaluation of Articulation and Phonology (DEAP; Dodd, Hua, Crosbie, Holm, & Ozanne, 2006), the Goldman-Fristoe Test of Articulation (GFTA-3; Goldman & Fristoe, 2015), and the Secord Contextual Articulation Tests (S-CAT; Secord & Shrine, 1997), one of which is the Contextual Probes of Articulation Competence (CPAC) with a Spanish language version called the CPAC-S (Goldstein & Iglesias, 2006). Most of these assessments are intended for children, and their focus is on L1 articulation and phonological attainment. However, they measure L1 phonemic accuracy very similarly to L2 phonemic accuracy via clinician transcriptions based on their judgment of the accuracy of the phoneme to the standard or acceptable phoneme.

Kirk and Vigeland (2014) conducted a psychometric review of six commonly-used phonological assessments: (a) Bankson-Bernthal Test of Phonology (Bankson & Berthal, 1990), (b) Clinical Assessment of Articulation and Phonology, 2nd ed. (Secord & Donahue, 2014), (c) Diagnostic Evaluation of Articulation and Phonology (Dodd, Hua, Crosbie, Holm, & Ozanne, 2006), (d) Hodson Assessment of Phonological Patters, 3rd. ed. (Hodson, 2004), (e) Khan-Lewis Phonological Analysis, 2nd ed. (Khan & Lewis, 2002), and (f) Structured Photographic Articulation Test II featuring Dudsberry (Dawson & Tattersall, 2001). Results showed that none of the tests provided sufficient opportunity to evaluate all 11 core phonological error patterns, including velar fronting, deaffrication, and gliding of lateral liquid /l/ and rhotic liquid /r/. However, the CAAP-2 and DEAP did provide at least four opportunities to demonstrate error patterns in final consonant deletion, initial cluster reduction, prevocalic voicing, postvocalic devoicing, palatal fronting, and stopping of fricatives and affricates (p. 373). Furthermore, reliability measures were conducted on these two tests, but neither obtained a coefficient alpha of at least .90 for internal consistency; nor did they provide adequate correlation coefficients for test-retest reliability, interrater reliability, or intrascorer reliability.

Succinctly put, phonemic accuracy is a construct that has been measured in both research studies and via published instruments, and while some scales have been used, this factor is more commonly measured in binary terms, as either correct or incorrect. Due to the existence of rater bias, it becomes necessary to clearly define what constitutes correct and incorrect phoneme utterances. It is also crucial that interrater reliability measures be taken in studies utilizing native-speaker rater judgements.

Summary of Literature Reviewed

This literature review has previewed at least five different important areas related to L2 pronunciation. First of all, phonological acquisition, in particular that of L2 acquisition, was discussed in terms of the various factors that have been shown to influence it. These factors were grouped into interspeaker and intraspeaker factors. Next, the relevance of the functional load was addressed as it relates to prioritizing phonemes for research purposes. The functional load was also shown to be as a valid source from which phonemic accuracy instruments might be built. Various types of speech elicitation tasks were then reviewed as to their efficiency at investigating phonemic accuracy. More specifically, those of word reading, word repetition, and picture naming were examined Finally, two L2 pronunciation principles were discussed, and the element of rater judgment was prevalent in the analyses. Findings indicated that due to the existence of rater bias, interrater reliability needed to be established.

CHAPTER THREE: METHODOLOGY

The methodology encompassing this study was structured within a quantitative research vein which aims to test causal hypotheses and potentially generalize results to a specific population. IRB approval was obtained (see Appendix A) for the following methodology used in the present study. Two research questions guided the research design and the selection of statistical analyses to be conducted.

Research Questions

1. Is there a statistically significant difference in L2 phonemic accuracy based on the speech elicitation methods of word reading, word repetition, and picture naming of high functional load initial phonemes found in high-frequency nouns occurring in American English?
2. Does native language moderate the effect of speech elicitations method on L2 phonemic accuracy of high functional load initial phonemes found in high-frequency nouns occurring in American English?

Research Design

This quasi-experimental study consisted of a cross-sectional repeated-measures design in which all participants were randomly assigned to one of two instrument versions, both containing exposure to all three elicitation methods. The elicitation methods rotated after each item, starting with word reading, followed by word repetition, and then picture naming, which continued in this order throughout the entire instrument. In addition, each instrument version was comprised of the same target nouns in a different order but with each noun represented in all three conditions to allow for within-subject comparative analyses. There was no control group, no pretest, and no intervention.

The dependent variable was phonemic accuracy, whereby each phoneme under investigation was measured on a dichotomous scale as either sounding like a SAE pronunciation or not. Research Question 1 investigated the effects of one independent variable, speech elicitation method, on the dependent variable. Speech elicitation method consisted of three conditions: word reading, word repetition, and picture naming, each utilized at least once for each target phoneme. Research Question 2 investigated the effect of speech elicitation method on phonemic accuracy scores but added the moderator variable of native languages as the between factor. Because the study sample involved multilingual participants, a number of different native languages were anticipated. Thus, to prevent having too few cases to generalize inside some L1s, only the most common L1s were utilized for the moderator variable.

Internal Validity

According to Shadish, Cook, and Campbell (2002), internal validity threats are "reasons to think that the relationship between A and B is not causal" (p. 54). Of the nine threats to internal validity they outlined, two of them were plausible threats in this study design. Among the threats, there was some missing data resulting from participants having various technical issues either not allowing them to record their audio or not allowing them to upload their audio files, so *attrition* threats were an issue. In fact, of the 93 potential participants, nine were absent, and only 61 produced audio files with no technical issues. Next, the *testing* threat to internal validity presented itself because participants were exposed to the conditions and target words multiple times; however, an attempt to reduce this effect was handled by changing the elicitation method frequently within the instrument.

External Validity

Shadish, Cook, and Campbell (2002) also outlined five threats to external validity, defining them as the reasons why the effects might not be generalizable to other populations, settings, treatments, and outcomes (p. 87). Because there was no random selection and no intervention, there were three threats to external validity in the present investigation. One threat, the *interaction of the causal relationship with units*, presented itself because results cannot be generalized to populations that were not sampled for the present study. Since mostly Chinese speakers and Arabic speakers were sampled, in the end, results cannot be generalized to Japanese speakers, who were not sampled. In addition, the program from which the participants in this study were sampled is unique and relatively new. Most English for academic purposes (EAP) programs are administered through community colleges and not within large state universities, so findings may not be generalizable to all EAP program participants but rather only to adult English learners in similar academic environments.

The threat of the *interaction of the causal relationship with settings* could also influence external validity due to the fact that the setting of the present study was conducted in groups of 10-15 participants at the same time in a laboratory on a large campus of a post-secondary institution. Had the data been gathered individually, not in a lab setting, or within another department on campus, there might have been different effects. Thus, the setting of the present investigation might vary naturally in ways that could influence the outcomes and overall generalizability to other settings.

Another threat to external validity is the *interaction of the causal relationship with outcomes*. The outcome of the present study was phonemic accuracy of onset sounds, so generalizations based on phonemic elements, such as final or medial position phonemes, cannot be derived here. This threat could also exist if a different instrument were used to measure phonemic accuracy. All in all, overall pronunciation accuracy and intelligibility cannot be generalized from the outcomes of the present study.

Participants

The sampling frame of the present investigation was adult English language learners over the age of 18 enrolled in pre-academic, undergraduate-track, English for Academic Purposes (EAP) courses at a large, public, post-secondary institution in the southeast region of the United States. The EAP Program offers two courses, EAP I and EAP II, each semester, with multiple sections of each; class size consists of approximately 15 students per section. During the data collecting phase of the present study, seven course sections were utilized, rendering a potential participant pool of 93. Study participants were recruited from all intact classes in the program, and participants were not obliged to participate; however, no attrition resulted from any participant opting not to participate.

The participants' English language proficiency was considered within an advanced range, based on program standardized test entry requirements. As described in Table 1, all EAP students should have TOEFL iBT scores of 60 or higher, with no individual section score below 12, or equivalent scores on other standardized tests.

Table 1

English Language Proficiency of Study Participants

Test	EAP 2 Requirements	EAP 1 Requirements
TOEFL	68-79 (No section score below 14)	60-79 (No section score below 12)
IELTS	5.5-6.0 (No section score below 5.0)	5.0-6.0 (No section score below 4.5)
Pearson Versant	50-68 (No section score below 42)	45-68 (No section score below 37)

Note. Adapted from Program website.

Finally, with regard to native languages, the L1 of the study groups was not monolingual because the EAP program admits student from all over the world, with a variety of different native languages. In the end, there were L1s of Chinese, Arabic, Urdu, Spanish, Vietnamese, and Russian, among others (see Figure 1). Therefore, the study groups were multilingual in nature and grouped based on English language proficiency level.

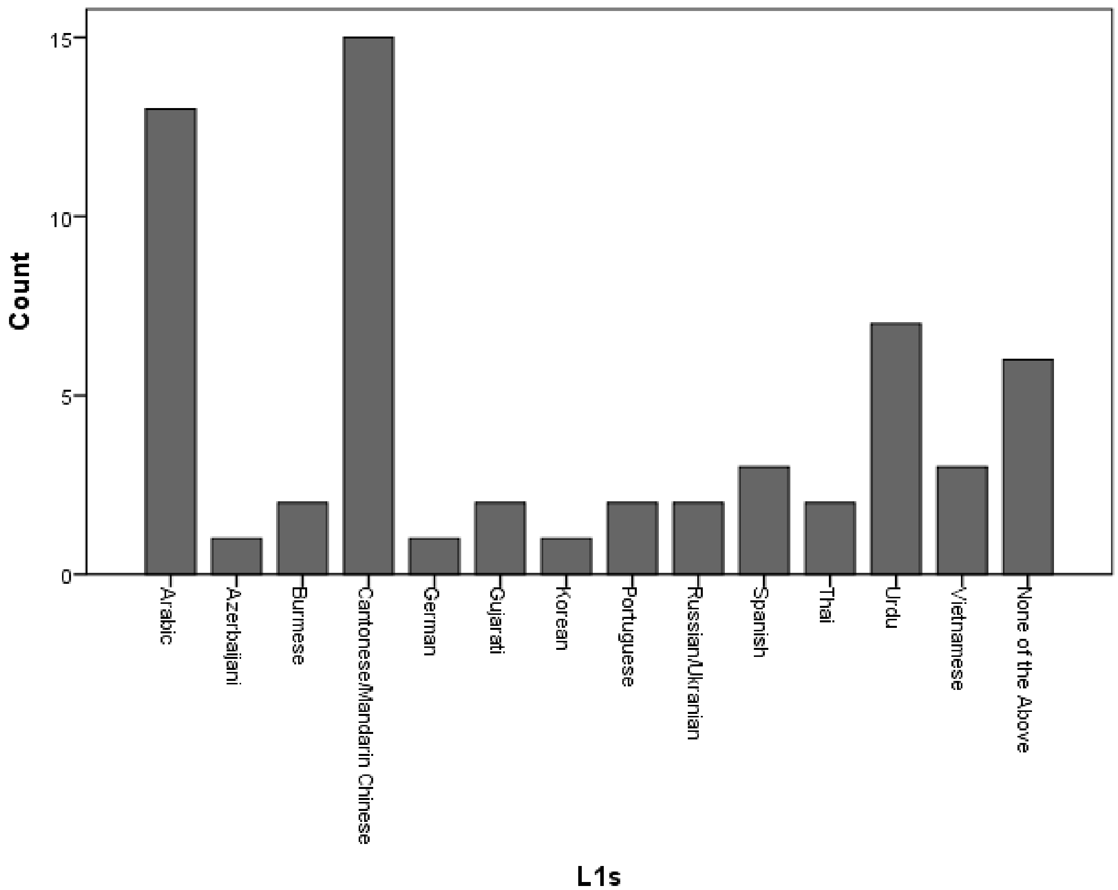


Figure 1. Participants' self-disclosed L1s.

Instrumentation

Data for this study was gathered through an instrument called the Speech Elicitation Method Instrument (SEMi), built around a corpus-informed word list consisting of high functional load phonemes found in the most frequently occurring picturable nouns. Studies analyzing L2 phonemic accuracy have commonly used word lists to elicit specific problematic phonemes. How these word lists have been compiled varies depending on the research questions under investigation. There are several well-known frequency-based word lists already available, including the General Service List (West, 1953), the University Word List (Xue & Nation, 1984), and the Academic Word List (Coxhead, 2000). More recently, Gardner and Davies (2014) devised a new word list called the Academic Vocabulary List (AVL), derived from the 520-million-word COCA (Davies, 2015). In addition, COCA offers free word frequency lists, one that includes the 5,000 most frequently occurring words in American English.

Therefore, for purposes of the present study, the SEMi was derived from corpora and using Nation and Webb's (2011) six steps to follow when compiling a word list: (1) decide on research questions or reasons for the list creation, (2) decide the unit of counting – word type, lemma, or word family, (3) select or create a suitable corpus, (4) decide on criteria for inclusion onto the word list, (5) decide on the criteria for ordering the words on the list, and (6) cross-check the created list against other corpora.

For the first step, the research questions of the current study involve examining frequently occurring American English words, so corpora of American English words were necessary to consult. In addition, the dependent variable, phonemic accuracy, can easily be measured via word lists, so the creation of a novel word list was required to accommodate all variables inside the research questions. For the second step, because analysis occurs on a phonemic level, single words were selected and not phrases, phrasal verbs, idioms or other concepts involving more than one word. Furthermore, the unit of counting was decided as word type due to the fact that nouns allow ease of elicitation through pictures, one of the elicitation methods under investigation.

For the third step, a couple of recommendations were considered when selecting the corpus to be utilized. The first was that the corpus include a minimum of one to three million words (Brysbaert & New, 2009; Coxhead, 2000), and another was that the corpus should include samples which are representative and generalizable to the linguistic environment under investigation (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Davies, 2015, Nation & Webb, 2011). The COCA (Davies, 2015) corpus not only has over 520 million words, but it also offers a representative sample of American English, the linguistic environment of the present study. Therefore, the COCA list of 5,000 most frequently occurring words was selected.

Nation and Webb's (2011) fourth step involves deciding which words to include in the instrument, a step which involved many filters before arriving at the items to comprise the SEMi. Because one of the elicitation methods is picture naming, only lexical items which can be represented pictorially were selected. Of all of the studies eliciting pictures, some used black and white images and others used color images. For the present study, color depictions were preferred and, thus, permission was secured (see Appendix B) to use the 360 color images from Moreno-Martinez and Montoro's (2012) study. The COCA nouns were cross-referenced with the 360 color pictures from Moreno-Martinez and Montoro's (2012) study, resulting in 124 picturable nouns.

The resulting consideration for inclusion onto the list involved determining initial functional load minimum requirements. Of the resulting 124 picturable nouns, only those with initial relative functional loads of more than 50%, per Catford (1987, p. 89-90), were selected. These sounds were treated individually in the present study rather than in minimal pairs as Catford had presented them, and in cases of duplicate phoneme representation, the highest relative functional load for each phoneme was assigned. Based on this criterion, the list of words was reduced to include 106 nouns with the following consonant onsets: /k/, /h/, /p/, /b/, /t/, /s/, /l/, /ɹ/, /b/, /d/, /f/, /w/, /g/, /n/, /m/, /ʃ/.

Monosyllabic nouns were selected in order to minimize interference from suprasegmental factors, and the resulting 76 monosyllabic items were further reduced to reflect only those with a high picture name agreement (.75), according to Moreno-Martinez and Montoro's (2012) study findings. Because the Moreno-Martinez and Montoro's (2012) study was norm-referenced for Spanish speakers, the English name agreement of the remaining 51 picturable nouns was then crossed with the Snodgrass and Vanderwart's (1980) study, from which Moreno-Martinez and Montoro based their expanded list. Any word not on the Snodgrass and Vanderwart list was removed. At this point, 36 picturable nouns remained, but six of the items were deemed ambiguous and were eliminated, and the word *gun* was removed for possible political and social implications, resulting in a total of 29 items.

The subsequent decision made for word inclusion involved identifying the initial phonemes of the remaining 29 words with the most occurrences to ensure relatively equal representation across the three elicitation methods. Each individual phoneme did not allow enough representations; some held counts as high as four, and others held counts as low as one. Therefore, individual phonemes were grouped by place and manner of articulation. Manner of articulation yielded less fragmented results with 12 stop onset items and 11 fricative onset items for a total of 23 usable items. The nasals and liquids held too few counts, so the decision was made to reserve those six words for the non-scored items.

Finally, because each item was elicited three times, the number of total items was reduced even further to prevent testing fatigue. The 23 items were reevaluated to strengthen the study instrument design by reflecting higher functional loads of at least .75 and picture name agreements of at least .8 in both the Moreno-Martinez and Montoro's (2012) and Snodgrass and

Vanderwart's (1980) studies. This gleaned 16 items, which were reduced once more to include no more than two occurrences of each phoneme. In cases where there were more than two items per phoneme, the two possessing the highest picture name agreement were retained; in addition, *train* was removed to better balance the ratio of stops to fricatives and because it was the only item without another item holding the same functional load percentage. This yielded the ultimate list of 12 items to be used for data collection purposes via the SEMi (see Table 2).

Table 2

Scoreable Items in the SEMi

Picturable Nouns	Functional Load %	Name Agreement % (Snodgrass & Vanderwart, 1980)	Name Agreement % (Moreno-Martinez & Montoro, 2012)
cat	100	100	100
cow	100	93	100
hand	100	93	100
horse	100	100	99
pen	98	95	100
ball	98	93	89
drum	82	98	92
duck	82	95	93
skirt	81	98	99
socks	81	100	100
foot	77	95	100
fork	77	100	100

For the non-scored items, the researcher wanted to have at least two items for each speech elicitation method to allow for acclimation toward each method, so the six nasal and liquid picturable nouns were set aside to be added later to the beginning of the instrument. For the tutorial, three more items were needed, one for each elicitation method, so the list of 51 words with high name agreement, per Moreno-Martinez and Montoro's was again reviewed, and three two words not already on the instrument were selected from that list. For the picture elicitation tutorial item, the word *chair* was added because it held the highest picture name agreement yet did not have a high functional load and had been eliminated earlier in the selection process. The final word count for inclusion in the instrument, therefore, was 12 items in addition to six non-scored items, and three tutorial items (see Table 3).

Table 3

Nonscored Items in the SEMi

Picturable Nouns	Functional Load %	Name Agreement % (Snodgrass & Vanderwart, 1980)	Name Agreement % (Moreno-Martinez & Montoro, 2012)
	Tutorial Items		
rose	83	N/A	98
wave	76	N/A	89
chair	39	100	100
	Nonscored Items		
lamp	83	93	100
leg	83	81	100
ring	83	98	82
nail	83	98	82
nose	61	98	100
moon	59	62	95

Nation and Webb's (2011) fifth step involves ordering the words, which in the present study consisted of assigning the words to the three speech elicitation methods: orthography/word reading (O), auditory/word repetition (A), and picture naming (P). In an attempt to avoid a testing effect which could be caused by grouping the speech elicitation methods together, the decision was made to rotate each elicitation method repeatedly over the entire instrument. The three methods for eliciting the target words were cycled in the following manner: the first item via word reading (orthography), the second item via word repetition (auditory), and the third item via picture naming (picture). This rotation of orthographic representation (O) followed by auditory representation (A) followed by picture representation (P) continued from the first to the last item, which in the end occurred 42 times starting with the six non-scored items followed by the 36 scored items (12 items across three elicitation methods).

As for ordering the 12 scored items within the already established order of the speech elicitation methods, an attempt was made to avoid the same phoneme from occurring sequentially. Therefore, the 12 scored items were first put in order based on initial phonemes with the highest functional load percentages first. Then the first four nouns on the list (*cat*, *cow*, *hand*, and *horse*) were assigned to the first four O positions on the instrument, so every third item. The second four nouns (*pen*, *ball*, *drum*, and *duck*) were assigned to the first four A positions on the instrument. The last four nouns (*skirt*, *socks*, *foot*, and *fork*) were assigned to the first four P positions on the instrument. To assure that all 12 nouns appeared with each of the three elicitation methods, this process of assigning the 12 words to the three SEMs was repeated two additional times, one for each of the other two SEM. The SEM assignment used in this instrument design protocol assured that all 12 words appeared once within the instrument in each of the three SEMs, hence eliciting 36 utterances of 12 words.

Two versions of the instrument were created to minimize a testing effect were the instruments to launch concurrently. The ordering of the first 12 items from SEMi A was replicated and placed as the second 12 items on SEMi B. This process was repeated for the second and third elicitation methods of the items so that no two words would ideally be elicited at the same time between the two instrument versions. In the end, both instrument versions contained an ordering of the 12 items in a relatively even distribution of high and low functional load across the two versions of the instrument, and both versions followed the same sequencing of speech elicitation method, only with different nouns (see Table 4).

Finally, the ordering of the non-scored items followed almost the same process, except that the four picture naming items were ensured to possess the highest name agreement of the six items. For the tutorial, as illustrated in Appendix C, the decision was made to use the same tutorial for both instrument versions. The noun *chair* was placed in the picture naming location due to its very high name agreement score, while *wave* was selected for auditory representation, and *rose* was selected for orthography representation.

Table 4

Word Order of SEMi A and SEMi B

Speech Elicitation Method	SEMi A	SEMi B
O	*ring**	*lamp**
A	*nose**	*ring**
P	*lamp**	*nose**
O	*nail**	*leg**
A	*moon**	*nail**
P	*leg**	*moon**
O	*cat*	*pen*
A	*pen*	*skirt*
P	*skirt*	*cat*
O	*cow*	*ball*
A	*ball*	*socks*
P	*socks*	*cow*
O	*hand*	*drum*
A	*drum*	*foot*
P	*foot*	*hand*
O	*horse*	*duck*
A	*duck*	*fork*
P	*fork*	*horse*
O	*skirt*	*cat*
A	*cat*	*pen*
P	*pen*	*skirt*
O	*socks*	*cow*
A	*cow*	*ball*
P	*ball*	*socks*
O	*foot*	*hand*
A	*hand*	*drum*
P	*drum*	*foot*
O	*fork*	*horse*
A	*horse*	*duck*
P	*duck*	*fork*
O	*pen*	*skirt*
A	*skirt*	*cat*
P	*cat*	*pen*
O	*ball*	*socks*
A	*socks*	*cow*
P	*cow*	*ball*
O	*drum*	*foot*
A	*foot*	*hand*
P	*hand*	*drum*
O	*duck*	*fork*
A	*fork*	*horse*
P	*horse*	*duck*

Note. O = Orthography, A = Auditory, P = Picture.

Data scoring. A rating spreadsheet was created with one sheet, or tab, for each instrument version with the 36 words in the order in which they were presented on SEMi A and SEMi B. In addition to the rating spreadsheet, a separate file was compiled containing all of the rating protocol. All deidentified audio files were kept intact for analysis. Raters advanced each audio track to about the four-minute mark to skip any audio captured from the tutorial and the first six words of the instrument, for those were considered warm-up items and were not scored. The rest of the words were analyzed for initial phoneme accuracy, defined dichotomously according to SAE. Working from only the unidentifiable audio files sorted into numerical order, the initial phoneme of each item was rated for accuracy by placing a "1" next to all accurate initial phoneme productions and a "0" next to all inaccurate initial phoneme productions. Part of the protocol depicted in Appendix D provided specific item-level guidelines for analyzing anomalies, such as the use of N/A (not applicable) for words that had recording issues, words that were not spoken, or completely different words that were uttered in place of the target word. These N/A items were treated as inaccuracies and later coded as "0." Each participant was later assigned four scores based on accurate responses: a word reading score, a word repetition score, a picture naming score, and a total accuracy score.

Data Collection Procedures

Upon IRB approval, the data collection process commenced. The study took place in a lab during 30 minutes of participants' regularly-scheduled classes. Some participant identifiable information was obtained from course instructors prior to data collection and kept in a secure location by the researcher. This information enabled adding the participants to the institutional online learning platform. Prior to data collection, 16 workstations located in the multimedia facility on the main campus were blocked off and designated as either A or B stations sequentially around the lab to prevent participants from sitting too close to other participants with the same test version. Data collection occurred during designated class times and was conducted over a one-week period to allow all classes to participate; recruitment occurred one week prior to data collection via course instructors informing their students of the in-class activity to take place. No payments or course credit was given as compensation for participation in the study.

The study was administered on the university's main campus in a multimedia facility, one equipped with computer workstations which allowed adequate spacing between study participants. Each computer had mid-tiered enterprise level towers with Core 2 Duo processors, professional 17-inch monitors, and Logitech H360 headsets with boom-type microphones, or similar hardware. Each computer was equipped with Voice Recorder, a recording device, Windows Operating system, and access to the online institutional learning platform. All of the hardware and software were individually tested for quality by the researcher prior to the execution of the study.

The entire data collection process lasted approximately 30 minutes. All of the directions for completion of the study were located inside the modules section of a secure online learning platform. Each module was labeled as the generic name of A and B and included one of the two versions of the instrument. Inside each module, there was a short tutorial video of instructions for completing the activity followed by the instrument, set to be viewed only after viewing the tutorial. Settings ensured that participants could not skip a step before launching the instrument.

Prior to entering the lab, participants were randomly assigned a number between one and 16, corresponding to a station number, previously identified to allow space between participants and minimal distraction. Once all participants were seated at a station, they were instructed briefly as to the purpose of the study and reminded to refrain from opening anything on their computer until instructed to do so. At this point, they were also explained about the option to participate via choosing whether or not to accept the course invitation that was sent soon thereafter.

First, the participants were asked to log into their stations using their institutional credentials. The researcher then showed them via a projector where they could find the recording device on their computer. Once all participants located the recording device, they were instructed to run an audio test of their equipment. Any necessary adjustments were handled immediately. During this time, invitations to join the secured website were sent to all participants.

After the audio equipment testing was complete, participants were instructed to log into the institutional online learning platform and asked to accept the website invitation. Accepting the invitation served as their consent process and as their waiver of written consent. No participant opted not to participate. Each station had an index card labeled “A” or “B” placed within view of the monitor. Once the participants logged into the learning platform and accepted the invitation, they were instructed to go to the module corresponding to the letter card of their station.

Each module consisted of content pages that had to be viewed in sequential order (see Appendix E for an example). The first page asked participants to answer a brief survey of demographic questions regarding their age and native language. The second page asked participants to initiate their recording devices; the researcher confirmed that all devices were properly recording. The third page introduced a three-minute tutorial video, consisting of three practice items, while the fourth page actually launched the tutorial experience. Within this video, participants were encouraged to practice speaking in a full voice as they experienced each of the sample items in the tutorial. They were told to speak each word only one time to practice. They were also encouraged to pronounce the words to the best of their ability and to attempt to imitate SAE pronunciation of the words they encountered. Immediately after the tutorial video was viewed, directions appeared on the screen explaining how to launch the SEMi.

The fifth page of each module launched either version A or B of the SEMi, the five-minute study instrument (see Appendices F and G). Upon completion of the instrument, the sixth page instructed participants to stop recording, thus allowing the files to automatically save to a designated folder within each participant's documents. The seventh page provided instructions on how to find the audio files and submit them to the assignment contained within the page. The final content page announced the end of the activity and indicated that participants could log off the computers.

After each group of study participants finished uploading their audio files to the proper location within the secured website, they were thanked for participating, reminded to log out of learning platform and off the station, and offered light refreshments as they exited the lab. The entire data collection process took about a half an hour. Once the lab was vacated, the researcher downloaded all audio files submitted onto a secured drive and deactivated all participants from the online learning platform course. All sound files were securely stored on the researcher's personal drive for later analysis.

Data Analysis Procedures

The first data set consisted of biographical information that was obtained from the secure online learning platform self-reported questionnaire on L1 and age. The second data set consisted of audio data. Both data sets were stored in a secure location at all times.

Data Preparation

In preparation of the data analysis phase, all audio data were re-labeled to remove identifying information. As one of the functions of the online learning platform, any uploaded file receives an automated name consisting of the participant's last and first name followed by two strings of numbers and the name of the file as the participant saved it. To ensure anonymity, the original files were copied and then renamed retaining only the second string of identifying numbers, thus removing the names of the study participants. Both the original audio files and the anonymously labeled audio files were stored in a secure location. A file was created that contained the demographic information for each participant, the original and the deidentified audio file names, and the randomly assigned instrument version. This file was saved to a secured drive and consulted as needed in the data analysis process.

Next, the researcher established the final audio data set by briefly listening to each file for audible responses and completion of the entire instrument. Where there were participant audio files with cases of inaudibility, missing sections, or files with no recording, these participants were removed from the final audio data set. The final audio data set (n = 61) was subdivided at this point in preparation for rating. Three audio data subsets were created to establish interrater reliability. The first two were used for rater training and calibration and the last one was used for the final interrater reliability test. The researcher first randomly selected 20% of the participants' deidentified audio files from each of the two versions of the SEMi and copied them to another folder labeled, "Interrater." Of the remaining data, two additional sets of 10% of the participants' deidentified audio files from each instrument version were randomly-chosen and copied to two folders labeled "Calibration 1" and "Calibration 2."

Interrater Reliability

For the purposes of interrater reliability measures, the researcher planned to include a second rater with TESOL experience. The first rating was conducted by the researcher on deidentified audio files for all participants. Once the researcher rated all of the data, the interrater process began. The interrater training consisted of an initial meeting with the researcher and the rater to discuss the protocol, which was followed by the release of the data set labeled Calibration 1. Once the second rater's scores were uploaded to a secure website, the researcher determined if there was interrater reliability and score consistency on the first calibration set by running scale-level analyses using the Pearson product-moment coefficient. Since additional practice was needed, the rater was issued the Calibration 2 data set and the interrater training process was repeated. Once a reliability coefficient of .7 or higher was established within the Calibration data sets, the secondary rater was then provided with the Interrater data set and appropriate rating form. After the second rater provided values, a Pearson r was conducted on the composite scores from all items, and a Kappa was conducted item-by-item to establish interrater correlation and interrater agreement before other statistical analyses were conducted.

Statistical Analysis

Before the first research question was addressed, a couple of preliminary analyses were planned. First, it was planned to analyze data for an effect of instrument version on the difference in phonemic accuracy between the three elicitation methods, so a mixed-design factorial analysis of variance (ANOVA) was conducted for this purpose. Second, it was planned to test the 12 items from the SEMi for internal consistency on each subscale using Cronbach's alpha.

In response to Research Question 1, a repeated measures ANOVA was run to determine if there was a difference in L2 phonemic accuracy based on speech elicitation method. In response to Research Question 2, a two-factor mixed design ANOVA was run to determine if native language moderated the effect of speech elicitation method on L2 phonemic accuracy. Statistical significance was defined on data analyses by an alpha level of less than .05.

CHAPTER FOUR: RESULTS

The statistical analyses applied to the data informing this study consisted of several procedures which were based on the dependent variable of phonemic accuracy, the independent variable of speech elicitation method, and the moderator variable of L1. Before interpreting the results of the main research questions, preliminary analyses were run to determine interrater reliability, interrater agreement, and the effect of instrument version.

Preliminary Results

In order to establish interrater reliability, a Pearson r was computed on the total composite scores of the two raters on the interrater set ($n = 13$), or 20% randomly selected from SEMi A and SEMi B in the total data collection set. The reliability coefficient ($r = .857$) indicated strong interrater reliability in raters' perceptions of phonemic accuracy.

To establish interrater agreement, a Kappa coefficient was calculated for each item on the SEMi. Of the 36 items, 12 items held perfect Kappa scores ($\kappa = 1.00$), the two items of *skirt* (P) and *cat* (P) held Kappa coefficients of .74 and .79 respectively, while only four items did not achieve high Kappa correlations ($\kappa < .70$): *cow* (O), *cat* (A), *pen* (O), and *skirt* (A). The remaining 18 items held constant values with no variability, resulting in no Kappa statistic computed. When this occurred, a proportion of agreement was calculated for the 13 cases (see Table 5).

Table 5

Interrater Item Agreement of the SEMi in Alphabetical Order by Elicitation Method

Item	*K*	Proportion of Agreement
	Word Reading	
ball	1.00	
cat	1.00	
cow	-.08	
drum	*	.92
duck	*	1.00
foot	*	1.00
fork	*	.92
hand	*	1.00
horse	1.00	
pen	.35	
skirt	*	1.00
socks	1.00	
	Word Repetition	
ball	*	1.00
cat	.63	
cow	*	.92
drum	*	1.00
duck	*	1.00
foot	*	1.00
fork	1.00	
hand	*	1.00
horse	*	1.00
pen	*	.92
skirt	.59	
socks	1.00	
	Picture Naming	
ball	*	1.00
cat	.77	
cow	*	.85
drum	1.00	
duck	1.00	
foot	1.00	
fork	1.00	
hand	*	.92
horse	1.000	
pen	*	.77
skirt	.74	
socks	1.00	

*No Kappa statistic because of lack of variability (see Proportion of Agreement).

To determine the reliability of the instrument, a mixed-design factorial ANOVA was conducted on the two instrument versions across the three elicitation methods. First, statistical assumptions for normality were tested via examination of the residuals for SEMi A and SEMi B. Review of the S-W test for normality for SEMi A (SW = .899, df = 35, p = .004) and SEMi B (SW = .838, df = 26, p = .001) showed statistically significant differences for normal distribution, indicating that the scores were not normally distributed in either instrument version.

In terms of skewness, SEMi A scores were skewed negatively (-.727) but within acceptable ranges for normality with the standard error of .398, while SEMi B scores were skewed negatively (-1.516) and not within acceptable ranges for normality with the standard error of .456. In terms of kurtosis, SEMi A scores held a kurtosis of -.413 with a standard error of .778, which were within acceptable ranges for normality, and SEMi B scores held a kurtosis of 2.443 with a standard error of .887, again within acceptable ranges for normality.

The second statistical assumption for mixed-design factorial ANOVA analyses is that of sphericity, and this assumption was not met due to significance found in Mauchly's test of sphericity (χ^2 = 26.438, Mauchly's W = .634, df = 2, p < .001); therefore, the results from Greenhouse-Geisser were interpreted. Finally, the fact that participants were spaced throughout the lab made it unlikely that they would hear responses from other participants, and random assignment of participants to each instrument version helped ensure that the assumption of independence was met. The ANOVA indicated that there was not a statistically significant interaction between the instrument versions and the speech elicitation methods (F(1.46, 86.38) = .138, p = .805, η^2 = .002), nor was there a main effect for the instrument version (F(1, 59) = .345,

p = .559, η^2 = .006). These findings make the two SEMi versions more reliable. Both of these results suggest that the scores given for each elicitation method are consistent across the two versions of the instrument. The means and standard deviations for the phonemic accuracy scores from each speech elicitation method on SEMi A and SEMi B are presented in Table 6.

Table 6

Means and Standard Deviations for SEMi A and SEMi B

	SEMi A (N = 35)		SEMi B (N = 26)	
Speech Elicitation Method	*M*	*SD*	*M*	*SD*
Word Reading	11.51	.658	11.54	.811
Word Repetition	11.46	.701	11.58	.857
Picture Naming	10.37	1.664	10.58	1.391

In addition, the 12 items from the SEMi were tested for internal consistency on each subscale using Cronbach's alpha. Results indicated that the items within each elicitation method subscale and test version did not possess high internal consistency as depicted in Table 7. These results could be attributed to a lack of variability in scores for certain items, which had to be excluded from the analysis.

Table 7

Internal Consistency on Subscales of SEMi A and SEMi B

	SEMi A	SEMi B
Speech Elicitation Method	α	α
Word Reading	-.066	.426
Word Repetition	.048	.564
Picture Naming	.639	.411

Note. α = Cronbach's α

To further examine this issue, an item difficulty analysis was conducted. This analysis revealed that the majority of items from SEMi A and SEMi B were relatively easy, particularly in the word reading and word repetition elicitation methods. In fact, for word reading, eight of the 12 items received 100% accuracy scores from all participants in at least one of the two versions, and the lowest percentage of accurate responses was still relatively high, that of 81% for the item *pen*. The rest of the items held 91% or higher accuracy percentages.

When examining item difficulty for word repetition, there were also eight items with 100% accuracy scores from all participants in at least one of the two versions, and the item with the lowest percentage of accurate responses for all participants was 74%, also for the item *pen*. The rest of the items elicited via word repetition held 81% or higher accuracy for all participants. Therefore, the item difficulty analysis for word reading and word repetition showed that items elicited in one of these ways for both SEMi A and SEMi B were not difficult.

However, for the picture naming elicitation method, only four of the 12 items received 100% accuracy scores for all participants in at least one of the two versions. The item *skirt* was considerably more difficult with only 51% of all participants correctly uttering this word. In addition, there were two other items that were relatively more difficult: *foot*, which was accurately pronounced by 69% of all participants, and *drum*, which was accurately pronounced by 73% of all participants. Table 8 shows the proportion of participants obtaining phonemic accuracy for each item inside each instrument version and within each speech elicitation method.

Table 8

Item Difficulty of SEMi A and SEMi B in Alphabetical Order by Elicitation Method

Item	SEMi A (*N* = 35) Difficulty	SEMi B (*N* = 26) Difficulty
	Word Reading	
ball	.94	.92
cat	.94	.96
cow	.97	1.00
drum	1.00	1.00
duck	1.00	1.00
foot	.97	1.00
fork	.94	1.00
hand	1.00	.92
horse	1.00	.96
pen	.86	.81
skirt	.97	.96
socks	.91	1.00
	Word Repetition	
ball	.94	.96
cat	.97	.96
cow	.97	1.00
drum	.89	1.00
duck	1.00	1.00
foot	1.00	1.00
fork	.97	.96
hand	1.00	1.00
horse	.97	1.00
pen	.74	.92
skirt	1.00	.81
socks	1.00	.96
	Picture Naming	
ball	.91	.96
cat	1.00	.92
cow	1.00	.92
drum	.74	.73
duck	.86	.96
foot	.69	.73
fork	.89	.81
hand	.97	1.00
horse	1.00	.92
pen	.94	.88
skirt	.51	.81
socks	.86	.92

Note. Difficulty based on proportion of correctly uttered initial phonemes

Inferential Results

The null hypothesis for the first research question was that there would be no difference in phonemic accuracy scores based on speech elicitation method, and the null hypothesis for the second research question stated that native languages would not moderate the effect of speech elicitation method on phonemic accuracy scores.

Research Question 1

A repeated measures ANOVA was run to examine the difference in L2 phonemic accuracy based on speech elicitation method. To ensure that the statistical assumptions were met, the residual errors were tested for normality, and between-group variability was tested for sphericity. The Shapiro-Wilk test was statistically significant for each speech elicitation method ($SW_{word\ reading}$ = .680, df = 61, p = .000; $SW_{word\ repetition}$ = .655, df = 61, p = .000; and $SW_{picture\ naming}$ = .860, df = 61, p = .000), indicating that the scores were not normally distributed. Scores for the three elicitation methods were all negatively skewed (-1.466 for word reading, -2.095 for word repetition, and -.959 for picture naming), as the skewness values were larger than two times the standard error (.306). Likewise, each subscale score was leptokurtic: word reading scores held a kurtosis of 1.666, word repetition scores held a kurtosis of 6.217, and picture naming scores held a kurtosis of .186. These kurtosis values were also not within acceptable normality ranges for word reading or word repetition due to the standard error being .604. Figure 2 visually demonstrates how the assumption of normality was violated for each elicitation method.

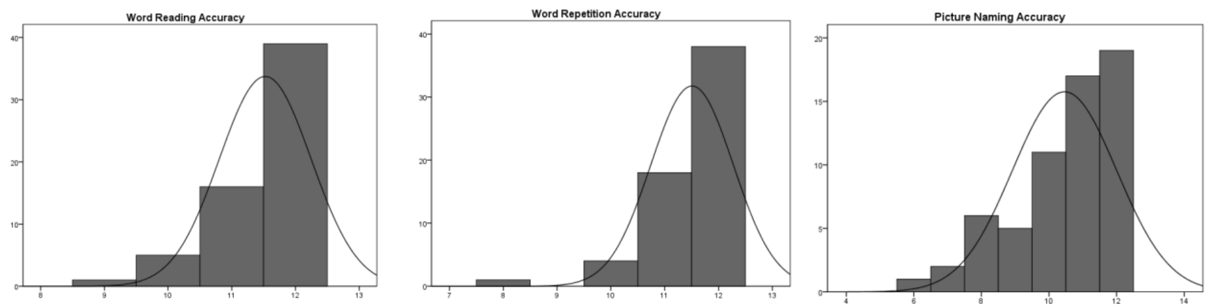


Figure 2. Histograms of word reading, word repetition, and picture naming scores.
X values are total scores, and Y values are total number of participants receiving each score

Since the assumption of normality was violated, in addition to the repeated measures ANOVA, Friedman's test was run to validate the results from the ANOVA. According to Mauchly's test, the assumption of sphericity was also not met ($\chi^2 = 26.770$, Mauchly's $W = .635$, $df = 2$, $p = .000$), so the results from Greenhouse-Geisser were interpreted in the repeated measures ANOVA. The assumption of independence within participants was controlled by spacing participants at least 10 feet apart from each other and by creating two instrument versions in alternating order around the lab to prevent any participant from sitting next to another participant with the same instrument version.

The results for the repeated measures ANOVA indicated a statistically significant effect ($F(1.47, 87.93) = 25.94$, $p < .001$) of speech elicitation scenario on L2 phonemic accuracy. The effect size was moderate (partial $\eta^2 = .302$), suggesting that approximately 30% of the variance in score could be accounted for by speech elicitation scenario, and observed power was high (1.00). The results from Friedman's test also indicated that there was a statistically significant difference in phonemic accuracy scores based on the speech elicitation method ($F_R(2, n = 61) =$

29.98, $p < .001$). The means and standard deviations for the phonemic accuracy scores from the repeated measures ANOVA and the mean ranks from Friedman's test are presented in Table 9.

Table 9

Means and Standard Deviations for Phonemic Accuracy Scores

Speech Elicitation Method	*M*	*SD*	*M Ranks*
Word Reading	11.52	.721	2.26
Word Repetition	11.51	.766	2.20
Picture Naming	10.46	1.545	1.54

Note. N = 61

Post-hoc pairwise analyses conducted with Bonferroni MCPs revealed that phonemic accuracy scores from picture naming differed significantly from both phonemic accuracy scores from word reading and phonemic accuracy scores from word repetition. Both tests indicated that the picture naming scores were significantly lower than the scores from word reading and word repetition, while phonemic accuracy scores from word reading and word repetition did not differ significantly. Results of post-hoc tests from both the repeated measures ANOVA and the Friedman's ANOVA are reported in Table 10.

Table 10

Pairwise Comparisons of Repeated Measures ANOVA and Friedman's ANOVA

	Repeated Measures ANOVA Post-Hoc Results (Bonferroni)			Friedman's ANOVA Post-Hoc Results (adjusted for Bonferroni)			
Pair	Mean Δ	*SE*	*p*	F_R	*SE*	*p*	$p_{adjusted}$
O–A	.016	.111	1.000	.066	.181	.717	1.00
O–P	1.066*	.179	.000	.721	.181	.000	.000
A–P	1.049*	.204	.000	.656	.181	.000	.001

Note. O = Orthography, A = Auditory, P = Picture.
* $p < .05$.

Therefore, the first null hypothesis that were would be no difference in L2 phonemic accuracy scores based on speech elicitation method was rejected at an alpha of .05 or lower.

Research Question 2

To answer Research Question 2, a two-factor mixed design ANOVA was run to determine if native language moderated the effect of speech elicitation method on phonemic accuracy. In order to avoid fragmentation of the data with only a few cases in some languages, only the most prevalent native languages were considered for this analysis, which were Chinese (n = 15) and Arabic (n = 13). Table 11 illustrates the frequencies of the participants' first languages.

Table 11

Distribution of Participant Native Languages

L1	*n*
Cantonese/Mandarin Chinese	15
Arabic	13
Urdu	7
Spanish	3
Vietnamese	3
Burmese	2
Gujarati	2
Portuguese	2
Russian/Ukrainian	2
Thai	2
Azerbaijani	1
German	1
Korean	1
None of the Above/Not Reported	7

The assumption of normality was tested via examination of the residuals for both Chinese and Arabic participants for each elicitation method. For the Chinese participants, review of the S-W test for normality ($SW_{WordReading}$ = .661, *df* = 15, *p* = .000; $SW_{WordRepetition}$ = .561, *df* = 15, *p* = .000; $SW_{PictureNaming}$ = .798, *df* = 15, *p* = .003) showed significance for all three elicitation methods, suggesting that the assumption of normality was violated for this group. Looking at the residuals for the Arabic participants, review of the S-W test for normality $SW_{WordReading}$ = .592, *df* = 13, *p* = .000; $SW_{WordRepetition}$ = .772, *df* = 13, *p* = .003; $SW_{PictureNaming}$ = .860, *df* = 13, *p* = .038) showed significance for all three elicitation methods, suggesting that the assumption of normality was also violated for this group.

A look at skewness and kurtosis values for each group can be viewed in Table 12. Chinese participants' scores on all three elicitation methods were negatively skewed. In terms of kurtosis, Chinese participants' word reading scores were not within acceptable ranges. Arabic participants fell within acceptable ranges for skewness and kurtosis for scores on each elicitation method. Even though the assumption of normality was violated for some L1 groups and some elicitation methods, ANOVAs tend to be robust against violations of normality.

Table 12

L1 Residuals for Each Elicitation Method

Speech Elicitation Method	Chinese (*n* = 15)				Arabic (*n* = 13)			
	Skewness	Std. Error	Kurtosis	Std. Error	Skewness	Std. Error	Kurtosis	Std. Error
Word Reading	-2.012	.580	4.867	1.121	-.946	.616	-1.339	1.191
Word Repetition	-1.176	.580	-.734	1.121	-.572	.616	-.332	1.191
Picture Naming	-1.388	.580	1.319	1.121	-1.258	.616	1.477	1.191

Since the assumption of sphericity was not met (χ^2 = 15.021, Mauchly's *W* = .548, *df* = 2, *p* = .001), Greenhouse-Geisser adjustments were used with these results too. ANOVA results showed that there was not a statistically significant interaction between speech elicitation method and native language ($F(1.40, 35.82) = .794$, $p = .417$, partial $\eta^2 = .030$), indicating that L1 did not moderate the relationship between speech elicitation method and phonemic accuracy. Tests of between-subjects effects revealed that L1 did not have a statistically significant effect on phonemic accuracy scores regardless of elicitation method ($F(1, 26) = .231$, $p = .635$, $\eta^2 = .006$). However, a statistically significant difference in phonemic accuracy scores was still found within participants based on speech elicitation method ($F(1.38, 35.82) = 13.14$, $p < .001$, partial $\eta^2 = .009$). The means and standard deviations of each L1 group can be seen in Table 13.

Table 13

Means and Standard Deviations for L1s by Speech Elicitation Method

	Chinese (*n* = 15)		Arabic (*n* = 13)	
Speech Elicitation Method	*M*	*SD*	*M*	*SD*
Word Reading	11.47	.834	11.69	.480
Word Repetition	11.73	.458	11.38	.650
Picture Naming	10.60	1.502	10.31	1.797

For a clearer visual representation of speech elicitation method scores by L1, profile plots were run, which are depicted in Figure 3 for the native language groups of Chinese and Arabic.

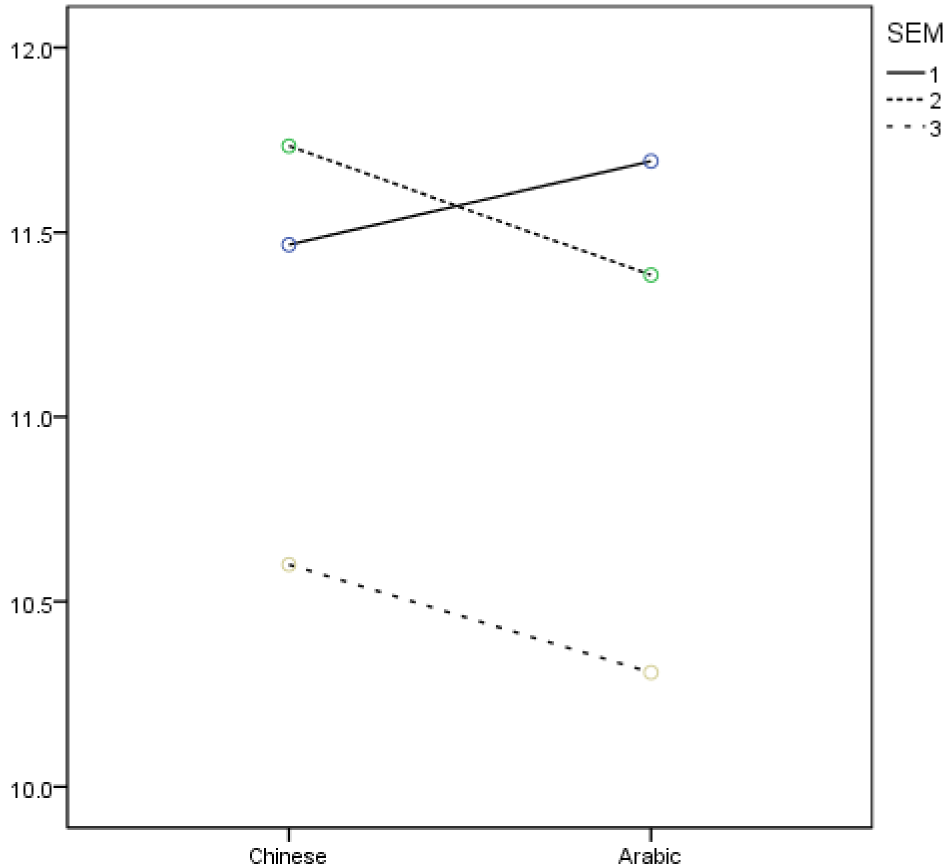


Figure 3. Profile plots for Chinese and Arabic participants across speech elicitation methods. Solid line = word reading, dotted line = word repetition, broken line = picture naming.

Therefore, the findings failed to reject the second null hypothesis that native languages would moderate the effect of speech elicitation method on L2 phonemic accuracy scores at an alpha of .05 or lower.

In sum, results for the first research question obtained from both the repeated measures ANOVA and the Friedman's ANOVA showed an effect of speech elicitation method on the L2 phonemic accuracy of high functional load initial phonemes found in high-frequency nouns occurring in American English. The difference was most prevalent in the picture naming elicitation method, while no statistically significant difference was found between the other two elicitation methods of word reading and word repetition. While there was a statistically significant difference in L2 phonemic accuracy scores based on speech elicitation methods, results for the second research question revealed that native language did not moderate the effect of speech elicitation method on L2 phonemic accuracy.

CHAPTER FIVE: CONCLUSION

Factors that are paramount to achieving L2 pronunciation accuracy in SLA are multiple and variable across and within learners. The present inquiry examined the effects of three different speech elicitation methods on L2 phonemic accuracy, namely word reading, word repetition, and picture naming. As Flege (2012) stressed, the role of input needs to be taken into consideration in future research on L2 pronunciation. There is a gap in empirical studies looking specifically at the three methods together examined within this study, and few, if any, have focused specifically on L2 phonemic accuracy based on the elicitation methods presented here within. All in all, it is important to highlight factors and limitations that surfaced in consequence to the methodology, with respect to each of the speech elicitation methods, and due to native language factors.

First, a discussion of statistical violations is needed. In each statistical analysis done, the assumptions of normality and sphericity were violated. An attempt was made to control for these violations by spacing participants apart in the lab and by randomly assigning them to one of the two instrument versions; however, these statistical violations may threaten the validity of the study findings. Indeed, phonemic accuracy scores were greatly skewed toward the higher end, so factors inherent to the instrument and data scoring were likely responsible for these statistical violations and the findings from each research question.

Instrument Discussion

The two versions of the SEMi allowed that each participant would experience all three conditions of word reading, word repetition, and picture naming for each item. An advantage of this within-subject design is that the participants serve as their own control group, but a limitation is that within-subject comparisons likely have carryover effects from one speech elicitation method to another. Therefore, participants' performance could have been decreased due to fatigue or increased because of practice or sensitization effects.

One could argue that the reason for such overall high phonemic accuracy across the tasks was related to the selection of only high-frequency words for inclusion in the SEMi, words that could have potentially been spoken by the NNSs on many different occasions. Trofimovich, Gatbonton, and Segalowitz (2007) found lexical frequency to be one factor determining the course of L2 phonological learning; thus, more frequent words were pronounced with greater accuracy in their study than less frequent words. Indeed, the item-difficulty analysis revealed in the present study that the initial phonemes of most items were pronounced accurately.

However, caution should be used when inferring that high-frequency words always result in higher phonemic accuracy. Indeed, what it means to "know a word" is multifaceted (Folse, 2004), and knowing a word may not equate to knowing how to pronounce the word with SAE accuracy. There are some words in a foreign language that are difficult for some NNSs to pronounce, regardless of their relative frequency. For example, the frequent word *water* can be difficult for NNSs of English to pronounce due to the presence of the /ɾ/ in SAE pronunciation. Even Trofimovich, Gatbonton, and Segalowitz (2007) found that much depends on the location of the phonemes within words and sentence contexts. Ultimately, whether words are frequently occurring or not, language learners utilize a numbers of communication strategies to avoid uttering those words that they find difficult to pronounce intelligibly.

In the present study, post-hoc analyses revealed that the statistically significant finding of the first research question was due to the third method of picture naming. This method produced lower phonemic accuracy within most participants and between most L1 groups and overall possessed higher variability as compared to the other elicitation methods. The selected pictures were analyzed prior to inclusion on the SEMi for having high name agreement in two separate studies (Moreno-Martinez & Montoro, 2012; Snodgrass & Vanderwart, 1980). The pictures for *drum* and *skirt* were problematic for the present study's participants, yet both words held high name agreement percentages per the Snodgrass and Vanderwart (1980) and Moreno-Martinez and Montoro (2012) studies (98% and 92% respectively for *drum*; 98% and 99% for *skirt*), yet neither of those studies were normed for NNSs.

Data Scoring Discussion

Perhaps data scoring might have impacted these statistical violations more than instrument factors. First, only initial phonemic segments were considered during the rating process, so L2 phonemic accuracy scores were higher than they would have been if all phonemic positions had been considered. For instance, any phonemic inaccuracies based on SAE that were found in medial or final positions of words elicited were not counted as inaccurate in the data analysis. To illustrated this with one of the SEMi items, if *cow* were pronounced as /kɔ/, it would have been rated as accurate based simply on the production of the initial sound. Vowel sounds and word endings tend to possess high variability in both standard and nonstandard English varieties, yet neither vowels, which are common in medial positions, nor final position consonants were investigated. In fact, the researcher noted many other cases containing medial and final sound inaccuracies per SAE, yet because the initial phoneme was produced accurately, a "perfect" score for that item was assigned. Due to this limitation, a ceiling effect was present in the L2 phonemic accuracy scores.

In addition, raters were instructed not to rate responses where participants either did not utter a word or uttered a completely different word than the intended word. These N/A items were coded as inaccurate responses in the data analyses conducted, maintaining a dichotomous scale but allowing these items to be set apart for further analysis. For example, if a participant uttered a synonym of the intended word, such as *pony* for *horse*, this utterance would have been coded as N/A and, thus, inaccurate. Scoring data on a dichotomous scale can restrict variability across scores. This lack of variability, combined with the fact that many items received perfect scores on the item difficulty analysis, resulted in some statistical analyses, such as Cronbach's alpha, from providing very useful information.

Indeed, this item-level guideline of assigning an N/A score could have been problematic when one considers what exactly constitutes a different word. Sometimes a mispronounced word could have been interpreted as a different word, such the utterance of *bowl* for the intended word *ball* or the utterance of *pan* for *pen*. Ultimately, these N/A scores influenced the results of the phonemic accuracy scores of many participants because each participant was assigned an accuracy score based on the total number of items on the SEMi, thus a total accuracy score out of 36 and out of 12 for each speech elicitation method. This scoring procedure allowed for instances where no word was uttered or a completely different word was uttered to be factored into the score, reducing the percentage of correctly uttered phonemes.

To account for this distinction that was made and to investigate the effect of these N/A scores on phonemic accuracy, new proportions of accuracy were calculated by removing N/A scores from the analysis and adjusting phonemic accuracy scores for only items with clearly correct or incorrect responses. Analyses of both research questions were then rerun on only attempts where phonemic accuracy could be ascertained. Results from these secondary analyses did not show statistically significant differences in either the first research question ($F(2, 120) = .133$, $p = .876$, partial $\eta^2 = .002$) or the second research question ($F(2, 51) = .136$, $p = .266$, partial $\eta^2 = .051$), but better informed the research and discussion.

With these secondary analyses, each item was analyzed for frequency of occurrence of incorrectly uttered initial phonemes, what could be argued were “true” errors. In fact, only four items, *pen* (11.5% average across the three elicitation methods), *cat* (9.8%), *ball* (3.8%), and *drum* (3.3%), were inaccurately pronounced by 3.3% or more of all participants. The researcher noted during data analysis that voicing and aspiration, or lack thereof, often resulted in inaccurate productions, and indeed these four items possessed initial stop sounds, half of which require aspiration. The remaining eight items were inaccurately pronounced by fewer than 3.3% of all participants. Of those eight items, four items were only pronounced inaccurately by 1.6% or fewer of all participants and only inside one of the three elicitation methods, and for these four items, the other two elicitation methods held 0% inaccuracies. The item *duck* held 0% inaccuracies for all three elicitation methods for all participants. However, in the actual data analysis *duck* was rated as N/A for 9.8% of the participants in the picture naming elicitation method only. Evidently considering N/A as inaccurate in the data scoring influenced the error counts and, thus, statistical significance found.

Speech Elicitation Method Discussion

Based on the difference between the two analyses for Research Question 1, it appears that speech elicitation methods do influence phonemic accuracy, but this difference is only apparent when either speech is not elicited at all or different words are elicited. It is critical, therefore, to examine cases where no words were uttered and cases where other words than the intended word were uttered. First, the inferences for the absence of sound are several. The pattern of a missing utterance was most frequently seen within the picture naming elicitation method. However, there were also several cases of missing utterances within the word repetition elicitation method. Where pictures were involved, it could be argued that some participants either did not know which word was intended by the picture or perhaps did not know the word in English, their L2. As previously mentioned, the pictures had not been normed for NNSs. This pattern occurred with the word *drum* for eight participants from SEMi A and five participants from SEMi B even if the word had been previously elicited in one or both of the other speech elicitation methods.

Where the word repetition elicitation method resulted in no utterance, which occurred in eight cases, it could be argued that sound is fleeting, so perhaps the word was not uttered due to inadequate processing time. Although each station was checked before gathering data, one might speculate as well that this phenomenon was attributable to a mechanical issue where a defective headset or computer audio setting might have made the auditory elicitation of some words more difficult to hear. It is important to note that these eight missing utterances all occurred near the beginning of each instrument version and none toward the end. While there were no evident patterns of specific words not spoken for the three missing utterances found within the word reading elicitation method, it occurred twice at the beginning of the instrument and once at the end. Finally, picture naming resulted in 27 cases of no utterance, and this pattern occurred just as much in the beginning of each instrument version as in the end. Therefore, acclimation to the instrument could have had more of an impact on missing utterances in both word reading and word repetition elicitations, but not for picture naming elicitations.

In addition to omitted utterances, there were cases where another word was uttered in place of the intended word. As with the missing utterances, the most frequent pattern of word substitutions was also found within the picture naming elicitation method. Further examination revealed this pattern most frequently with the items *skirt*, *foot*, and *duck* regardless of their location on each instrument version. Commonly spoken substitutions for these words were *dress*, *leg*, and *bird*. Uttering a different word was not as common with the word reading and word repetition elicitation methods, but when it occurred with word reading, initial phonemes were substituted with initial cluster sounds resulting in different words being spoken (i.e. *scores* for *socks* and *flock* for *fork*). It only occurred in one case with word repetition (*hat* for *pen*) and this occurring in the beginning of the instrument, perhaps again the consequence of the fleeting nature of sound or not having yet acclimated to the instrument.

Clearly further analysis of the effect of speech elicitation method on the absence of sound and on the production of a completely different word is required. In the present study, picture naming resulted in highly irregular responses. This finding reflected the well-known expression that "a picture is worth a thousand words," or in the very least can conjure multiple words. Though one could argue not as likely to happen, even a native speaker could potentially replace a picture intended to elicit the word *foot* as *toe*. Much depends on the image itself, which is why the name agreement of each word is crucial to consider for future research.

Finally, while there were no statistically significant differences found between the phonemic accuracy of utterances derived from word reading and word repetition, one cannot forget that there is always a potential for an orthographic effect on phonemic accuracy. Research has shown that reading acquisition occurs faster in languages exhibiting more shallow orthographies as compared to languages with deeper orthographies (Frith, Wimmer, & Landerl, 1998; Koda, 2016; Landerl, 2000; Wang & Koda, 2007). Other studies have shown effects of response times and naming accuracies based on L1 (Akamatsu, 1999; Wang & Koda, 2007), such that NNSs of languages with less orthographic depth, such as Spanish which is highly phonetic, struggle more with the pronunciation of languages with more orthographic depth, such as English. In the present study, however, no statistically significant effects were found between phonemic accuracy scores from word reading (M = 11.54, SD = .67) and those from word repetition (M = 11.51, SD = .77). It seems that orthographic depth is more complex and may not necessarily impact the more specific focus of phonemic accuracy in word reading tasks.

Native Language Discussion

The number of cases per native language resulted in only the more prevalent L1s counted in the analysis, which reduced the total number of participants included in the second research question. Because the particular L1s of the study participants could not be firmly identified prior to the study, data on native languages was collected via self-reported measures whereby participants were asked to select one language from a list of many L1s. Self-disclosed ways of gathering data tends to be less reliable, and indeed there were seven cases of participants indicating "None of the Above" as their native language.

Some of the participants are multilingual speakers and might have been forced to select one language over another language even though they might be equally proficient in another language on the list. There were several participants, for instance, who indicated Cantonese as their native language but who are also fluent speakers of Mandarin Chinese. Other participants from China indicated "None of the Above" as their native language perhaps due to the fact that they were not able to discern Mandarin over Cantonese as dominant. All in all, the total number of participants ultimately included in the analysis of the second research question was influenced and reduced because of this issue.

The results of the second research question, which asked whether native language moderated the relationship between speech elicitation method and phonemic accuracy, indicated that for the two native languages of Chinese and Arabic there was not a statistically significant interaction. One can note considerable differences between these two languages. Chinese is a logographic language, whereby the written language is expressed via one character to represent a word or phrase; Arabic, on the other hand, is a phonetic language much like English but utilizing a different alphabet. In terms of orthographic depth, Chinese has a deeper orthographic depth than Arabic (Jiang, 2016), and cross-linguistic research has shown that readers with related L1 and L2 orthographic depth backgrounds tend to be at an advantage in L2 word recognition over those with less related orthographic backgrounds (Akamatsu, 1999; Wang & Koda, 2007).

However, perhaps the two languages share more similarities than one might expect, particularly in relation to their differences to English. For one, neither Chinese nor Arabic share as many cognates with English as one might find in other Romance languages, and as Hoshino and Kroll (2008) found in cross-language picture naming, the speed at naming pictures was facilitated by cognates in English. Moreover, both Chinese and Arabic utilize different writing scripts than the Latin alphabet of English. Traditional Chinese is written vertically, and Arabic is written from right to left, so both share different directionality from English. Chinese and Arabic script differences, including directionality, could have likely influenced performance for word reading, yet this shared characteristic between the two groups did not reveal significant findings.

While the focus of the present study did not examine specific phonemes for their accuracy within participants and across groups, one interesting observation relative to L1 was that the SEMi included several instances of initial /p/, which Arabic speakers tend to have difficulty pronouncing, yet results still showed a lack of significance for L1. Looking specifically at the scores for the two groups across the different speech elicitation methods, there were no significant findings, yet somewhat surprisingly, Arabic speakers held slightly higher word reading phonemic accuracy scores on average than Chinese speakers. Considering Arabic orthography is not as deep as that of English (Jiang 2016), one might expect lower results on word reading tasks from Arabic participants compared to Chinese participants who share deeper orthographies with English, but this was not found to be true for this study. Perhaps this was a function of the SEMi words selected, the choice of only initial phonemes for analysis, or the proficiency level of the participants. Orthographic depth can influence L2 reading acquisition, yet when the mere task of word reading is involved, perhaps it does not play as vital a role.

Implications

This study has several implications for L2 research, teaching, and learning. First, inquiry into speech elicitation methods in SLA is just emerging, so the present study further explored this area to see whether elicitation methods impacted L2 pronunciation accuracy. Next, the study added one more dot to the pixilation of research examining the functional load as a way of prioritizing phonemic elements of a language (Kang & Moran, 2014; Munro & Derwing, 2006). Finally, this study reflected the findings of past studies that have shown that perhaps L1 is not as important of a factor as it was once thought to be (Elliott, 1995; Thompson, 1991). While accuracy here was defined in terms of initial phonemic accuracy, implications can be made from the results of this study that can better inform the field of SLA and teaching and learning.

Research implications showed that the impact of elicitation methods may very well be related more to cognitive factors. In other words, the elicitation method influenced the words uttered and not necessarily the phonemic accuracy of those words. The findings of the first research question indicated that pictures may not be as reliable a measure for eliciting phonemic accuracy due to the added semantic component present in pictures. As research in cognitive psychology has shown, processing times are faster for semantic processing compared to phonological processing (Indefrey & Levelt, 2004; Schmitt, Münte, & Kutas, 2000). Because phonological code retrieval takes longer than lexical access retrieval, adding this semantic layer to the input modality could influence phonemic accuracy. It appears to be more efficient and result in better word accuracy and phonemic accuracy to use one of the other two elicitation methods of the present study, word reading or word repetition.

Another implication for research is related to the SEMi, the present study's measurement tool. The process by which the SEMi was created can impact future research on speech elicitation methods and L2 phonemic accuracy. While the SEMi is not yet validated for NNSs, it could contribute to the creation of a viable tool for measuring pronunciation accuracy. Pronunciation-specific assessments are needed in the TESOL field. This instrument could have important implications for placement into language programs, tests of pronunciation comprehensibility and accuracy in the language classroom, and even for employment purposes. Validating such an instrument would make an indelible mark on the SLA field.

Pedagogical implications are also plentiful. Many L2 teachers have indicated their relative lack of formal pronunciation training and preparedness with regard to teaching pronunciation (Grant, 2014), so from this research they can learn about various choices they have to elicit pronunciation that is more accurate from their students. Brought to light in the present investigation were several possible elicitation methods that teachers could implement when practicing and reinforcing pronunciation inside their classrooms. Practicing specific segments with pictures, audio, and orthographic representations could make a pronunciation lesson richer and more thorough. Regardless of the elicitation method used, L2 teachers could practice and reinforce aspiration and voicing of initial sounds with their students as these issues impacted phonemic accuracy in the present study. Ultimately, it is hoped that this study will help teachers feel more prepared to teach L2 pronunciation in their classrooms, something that students desire.

This study has clear implications for assessment and teaching of ELs. Even though the present study found picture naming tasks resulted in less accurate responses, pictures and picture naming could still have a place in the L2 classroom and for assessments. Pictures are not utilized in TESOL as much with adults; however, images and even picture naming are used considerably with children and even more so in assessments of speech sound disorders. Furthermore, with EL adults, pictures are incorporated in high-stakes testing, such as the TOEFL, yet their focus tends to be more on communicative tasks. Describing an image or narrating a story based on a series of pictures is commonly used in assessments of adults learning another language. Based on the findings of the present study, it seems to be better to assess more suprasegmental pronunciation features like stress, intonation, and fluency with these types of assessments rather than focus only on segmental accuracy of certain words. Because pictures result in variable responses, allowing multiple "correct" answers should be an important consideration for L2 teachers.

In K-12 classrooms, students are often provided with computers and headsets and asked to listen and repeat sounds, so there is a strong interest in modalities in the classroom, particularly considering the impact that technology is having on our current educational systems. There is also an emphasis on visual representation of instruction and assessment; however, visual representation alone can be subjective. Based on the findings of the present inquiry, educators have options and would be wise to include all methods of eliciting speech within their classes, keeping in mind that pictures tend to be irregular in their accuracy. Considering students have various learning preferences, including more rather than fewer manners of eliciting speech might be a smarter option in the classroom.

Language learners often learn new vocabulary via word lists, and in the beginning levels, pictures are commonly included alongside the orthography. Language teachers should also supplement any new word lists with word repetition, and this word repetition could potentially include native and nonnative pronunciation varieties to allow learners to differentiate subtle pronunciation differences. A reading teacher, for instance, should include visual and auditory methods of presenting material and have students practice with each elicitation method to allow students to perform under a variety of elicitation methods on future assessments; all in all, both instruction and assessment should be reflective of multiple elicitation types.

Next, while it is evident that learner differences exist based on L1s, teachers should not place too much emphasis on native language as a deciding factor over their students' pronunciation accuracy and encourage their students to do the same. The findings of this study did not find statistically significant differences in phonemic accuracy as moderated by L1, so there are other factors that seem to be influencing L2 phonemic accuracy more than native language. Moreover, teachers should encourage their students not to let their native language determine their ability to speak with more native-like accuracy.

Finally, implications of the findings from this research study can benefit L2 learners and NNSs in several ways. Students in second or foreign language classrooms can be better informed about the various ways in which speech can be elicited. Because neither word reading nor word repetition showed statistically significant differences, L2 learners can practice with both, and they can also become more cognizant of their own preferences and which method may serve their purposes best. An implication of the findings from the second research question for L2 learners, in general, is that their native languages may not actually inhibit their pronunciation accuracy. NNSs of other languages can be more confident in their ability to improve in their L2 productive accuracy and be more encouraged by other factors that are within their control to help them achieve better results.

Future Research

There is a need for further inquiry into elicitation methods and their impact on SLA. The effects of elicitation methods could be investigated in other productive skills, such as writing. Speech elicitation methods could influence more than L2 phonemic accuracy, so this realm of research should be expanded to include areas like intelligibility and comprehensibility. All in all, more studies looking into the effects of speech elicitation methods on language learning are in order.

First, it would be interesting to examine these elicitation methods in combination with each other to see if perhaps the modality of picture naming would be supported if accompanied by audio or orthography. Another option could be to combine the auditory elicitation with orthography to determine if having both present influences the accuracy results. While processing time was not examined in the present study, it was noticed that word reading tended to result in faster response times, so this factor could be investigated in the future with each elicitation method and the combination of elicitation methods. This study could provide richer information to L2 instructors on better ways to elicit more accurate speech from their students and if one or a combination of elicitation methods is more helpful.

Future studies could also examine this phenomenon by replicating the study design with low-frequency words to see whether these words reveal a stronger effect of speech elicitation method on L2 phonemic accuracy. Word frequency could be incorporated as one of the variables in future investigations by comparing both high-frequency words and low-frequency words within the participants. Much knowledge on the impact of word frequency and L2 pronunciation accuracy can be gained from analyzing this factor in more depth.

Another possibility could be to explore the same three speech elicitation methods but with different phoneme frequencies. The functional load offers an effective way to classify initial phonemes into relative high and low frequencies and, as such, should be utilized in more L2 research. It would be fruitful to examine low-frequency phonemes across the three elicitation methods to see if findings are different. The functional load could be utilized again in the development of a new word list, this time with only low functional load initial phonemes. Future studies could also include both high and low functional load initial phonemes to verify whether L2 phonemic accuracy differs significantly between the two groups.

In addition, the functional load involves more than just initial phonemic contrasts and frequencies. Future inquiry could focus on L2 phonemic accuracy in medial or final positions of the SEMi to see whether results reveal less accurate productions. Medial positions often involve vowels, which are quite variable in their acceptable pronunciation accuracies even within NSs, so a thorough look at what constitutes vowel accuracy is warranted and would be informative for future research efforts. Final positions would also require some investigation as to what constitutes accuracy as some languages and even varieties within languages enunciate word endings differently. Finally, once both medial and final phonemic accuracy is well defined, it would also be worthwhile to investigate all phonemes elicited within the SEMi to compare accuracy on initial, medial, and final positions.

Future research on speech elicitation methods in L2 pronunciation should be expanded upon to include more L1s. It would be interested to investigate the effect of speech elicitation methods with the SEMi but utilizing participants from other syllabic languages or phonetic languages, such as Romance languages that share more commonalities with English. Also worthy of further investigation is a replication of the study design with other languages to see whether results are inherent to English or can be seen in other second or foreign languages. All in all, speech elicitation methods could be examined using a similar methodological model as the present study but with other foreign and other L2 learning environments to see whether their findings replicate those of the present study.

When investigating the effect of L1 on future research findings, it is essential to find reliable ways to determine the native language of participants. If self-reported measures are the only viable option, then one suggestion for future investigators would be to remove the option for participants to indicate "None of the Above" as their native language. Nonetheless, identifying L1 in ways that are not self-reported might provide more reliable answers to research questions looking at L1 as a factor. In addition, L1s could be clustered differently in future data analyses, such as organizing them into phonetic, syllabic, or logographic groups or comparing within one group, like all Romance languages. Much can be interpreted based on L1s, so not only defining how they are collected, but also reorganizing points of comparison in future studies would be a worthwhile endeavor.

There was no native speaker control group in the present study, which prohibited verification that results were actually a consequence of SLA. Perhaps any observed effects were merely cognitive in nature and not an SLA issue. Future research, therefore, should add NS controls to its design in order to separate results that could have been merely cognitive from those that were a consequence of SLA. Native speakers can add depth to the findings of studies conducted with only nonnative speakers. For instance, by comparing across NS and NNS groups, one might resolve the issue with missing utterances and focus more on potential cognitive differences.

Regardless of the relative frequency of the words on the instrument, the relative frequency of the phonemes included in the analysis, or whether other second or foreign languages are included, it is vital to have an instrument that has been validated for NNSs. If picture naming is included, it is essential that each image hold high name agreement for NNSs.

A psychometric study validating a speech elicitation method instrument for NNSs would be useful to have for future pronunciation research and assessment in the SLA field. Such an instrument and future instruments developed thereafter could be instrumental in L2 pronunciation placement within language learning facilities. It could also provide opportunities for teachers to measure specific phonemes in the classroom or replicate the elicitation methods better suited for a particular population. Lastly, there is a need for a validated L2 pronunciation instrument in the TESOL field, one that could potentially evaluate international teaching assistants and international candidates hoping to secure employment at colleges and universities.

Finally, perhaps the reason for overall high phonemic accuracy is related to another factor not investigated in the present study, that of language proficiency. Based on some of the models proposed for L2 phonological learning (Flege, 1995; Gatbonton, 1975, 1978; Major, 1987, 2002), variations in L2 phonological forms that reflect L1 occur more in the beginning levels of proficiency in the L2. Language proficiency should be incorporated into future studies because it has shown to be a predictor of sound perception and pronunciation accuracy in previous studies (Bley-Vroman & Chaudron, 1994; van Els & de Bot, 1987).

To begin with, language proficiency could be determined via standardized placement tests, such as the TOEFL and IELTS. Once accurately identified for each participant, proficiency scores could be sorted into various points on an interval scale to designate levels and to allow for variability in the data analysis phase. Additionally, individual section scores could be crossed with the data to see whether they are correlated with scores within specific speech elicitation methods. For instance, it would be interesting to examine if higher reading proficiency scores correlates with better accuracy in the word reading elicitation method under investigation. Scores from the productive skills of speaking and writing could also be compared with scores from the receptive skills of listening and speaking.

Further analyses could be conducted that look at language proficiency crossed with L1. Future research findings would be enriched by a cross comparison of L1s with overall language proficiency and specific skill proficiency to see whether any trends emerge. Potential discoveries resulting from these factors and how they interact with each other would strengthen the design of any future study and add to the body of research investigating both L1 and language proficiency together. All in all, language proficiency is an important factor to consider in L2 pronunciation studies examining the effects of elicitation methods on phonemic accuracy.

Summary

The idea that input modalities impact SLA is a recently emerging perspective, and the growing awareness that speech elicitation methods have on L2 pronunciation accuracy has a vital role in this discussion. There are cognitive factors at play in the processing of words that are written, heard, or seen in pictures that have an influence on the accuracy of what is ultimately uttered, so evidently a lot is happening when language learners are asked to accurately pronounce words in the language.

Overall, the findings from each statistical analysis suggested that speech elicitation method has an impact on L2 phonemic accuracy; in particular, caution should be used when including picture naming as a way of evaluating phonological production. In addition, L1 might not be as important a factor when it relates to phonemic accuracy. It is hoped that this study and future ones continue to uncover these factors to better inform L2 researchers, teachers, and perhaps most importantly, the learners themselves.

APPENDIX A: IRB LETTER OF APPROVAL

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Human Research

From: **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To: **Nicole Hammond Carrasquel**

Date: **January 10, 2017**

Dear Researcher:

On 01/10/2017 the IRB approved the following human participant research until 01/09/2018 inclusive:

Type of Review:	UCF Initial Review Submission Form Expedited Review
Project Title:	Phonemic Accuracy of American English Words Based on Speech Elicitation Method
Investigator:	Nicole Hammond Carrasquel
IRB Number:	SBE-16-12833
Funding Agency:	
Grant Title:	
Research ID:	N/A

The scientific merit of the research was considered during the IRB review. The Continuing Review Application must be submitted 30days prior to the expiration date for studies that were previously expedited, and 60 days prior to the expiration date for research that was previously reviewed at a convened meeting. Do not make changes to the study (i.e., protocol, methodology, consent form, personnel, site, etc.) before obtaining IRB approval. A Modification Form **cannot** be used to extend the approval period of a study. All forms may be completed and submitted online at https://iris.research.ucf.edu .

If continuing review approval is not granted before the expiration date of 01/09/2018, approval of this research expires on that date. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

Use of the approved, stamped consent document(s) is required. The new form supersedes all previous versions, which are now invalid for further use. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Participants or their representatives must receive a copy of the consent form(s).

All data, including signed consent forms if applicable, must be retained and secured per protocol for a minimum of five years (six if HIPAA applies) past the completion of this research. Any links to the identification of participants should be maintained and secured per protocol. Additional requirements may be imposed by your funding agency, your department, or other entities. Access to data is limited to authorized individuals listed as key study personnel.

In the conduct of this research, you are responsible to follow the requirements of the Investigator Manual.

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:

Signature applied by Gillian Amy Mary Morien on 01/10/2017 05:50:51 PM EST

IRB Coordinator

APPENDIX B: PERMISSION LETTER FOR IMAGE USE

From: FRANCISCO JAVIER MORENO MARTINEZ [fjmoreno@psi.uned.es]
Sent: Saturday, August 20, 2016 4:59 AM
To: Nicole Carrasquel
Cc: FRANCISCO JAVIER MORENO MARTINEZ; Pedro R. Montoro
Subject: Re: permission to use images

Dear Nicole,

Thank you very much for your knid message and your interest in our work. Please, feel free to use our images in your interesting work and thanks again for your consideration.

Let us know if you need anything more from us.

Best regards.
F. Javier Moreno.

---------- Mensaje reenviado ----------

De: "Nicole Carrasquel" <Nicole.Carrasquel@ucf.edu>
Fecha: 19/08/2016 21:23
Asunto: permission to use images
Para: "prmontoro@psi.uned.es" <prmontoro@psi.uned.es>
Cc:

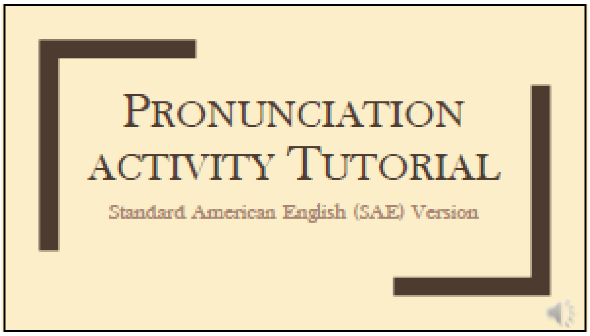
Dear Mr. Montoro,

My name is Nicole Hammond Carrasquel, and I working toward my PhD in TESOL at a university in the United States. My potential dissertation study will involve creating an instrument to elicit non-native English speech and investigate phonemic accuracy of picturable nouns. In my research, I came across your article, published with Mr. Moreno-Martinez, entitled "An Ecological Alternative to Snodgrass & Vanderwart: 360 High Quality Colour Images with Norms for Seven Psycholinguistic Variables. I am in the process of narrowing down my word list of frequently-occurring spoken content nouns, and I was hoping that I could use some of these 360 color images in the building of my instrument. If you give me permission to use these images, I would of course credit your invaluable research in my dissertation study. I hope it will serve as an instrument to assist in second language acquisition pronunciation research and instruction. If you would like more information about how I intend to use these images in my instrument, I would be happy to share more with you.

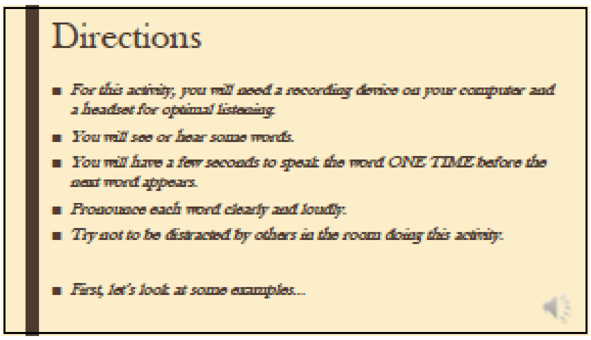
Thank you for your time and consideration!

Nicole Hammond Carrasquel, PhD Candidate
Visiting Instructor of TESOL
Department of Modern Languages & Literatures
University of Central Florida
Colbourn Hall 509

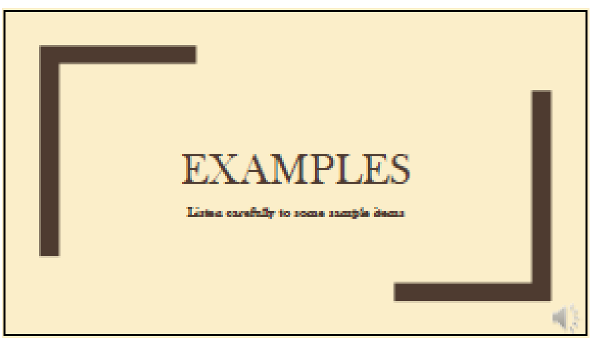
APPENDIX C: SEMI, TUTORIAL AND SCRIPT


PRONUNCIATION ACTIVITY TUTORIAL
Standard American English (SAE) Version

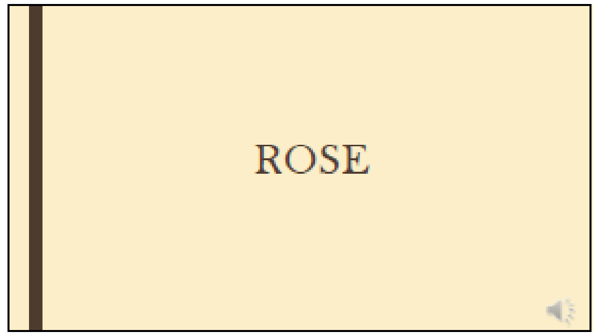
1


Directions
For this activity, you will need a recording device on your computer and a headset for optimal listening.
You will see or hear some words.
You will have a few seconds to speak the word ONE TIME before the next word appears.
Pronounce each word clearly and loudly.
Try not to be distracted by others in the room doing this activity.
First, let's look at some examples...

2


EXAMPLES
Listen carefully to some sample items

3


ROSE

4


[WAVE]

5

6


LET'S BEGIN
Remember to speak each word one time as clearly and as accurately as you can

7

Script for Tutorial Directions

Audio File 1 (Slide 1)…

Thank you for participating in this pronunciation activity. By now, you should have initiated the recording device on your computer. In this short tutorial video, you will receive directions for how to complete the activity. Please watch the entire video and practice speaking the sample items out loud before starting the activity.

Audio File 2 (Slide 2)...

Directions:

- *For this activity, you will only need a recording device on your computer and a headset for optimal listening.*
- *You will see or hear some words.*
- *You will have a few seconds to speak the word ONE TIME before the next word appears.*
- *Pronounce each word clearly and loudly.*
- *Try not to be distracted by others in the room doing this activity.*
- *First, let's look at some examples…*

Audio File 3 (Slide 3)...

Some sample items appear on the next three slides. Listen carefully to each item, and practice speaking each word out loud ONE TIME after you see or hear it.

Audio File 4 (Slide 4)...

In this example, you see the word "ROSE" written on the page. You will have a few seconds to speak the word "ROSE" out loud one time. Go ahead and speak the word out loud now to practice.

Audio File 5 (Slide 5)...

In this example, you will hear a word. (WAVE). You just heard the word "WAVE." You will have a few seconds to repeat the word "WAVE" out loud one time. Go ahead and speak the word out loud now to practice.

Audio File 6 (Slide 6)...

In this example, you see a picture of a "CHAIR" on the page. You will have a few seconds to speak the word "CHAIR" out loud one time. Go ahead and speak the word out loud now to practice.

Audio File 7 (Slide 7)...

This concludes the example items. Remember to speak each word one time as clearly and as accurately as you can…Now, click the "Next" button on your computer to initiate the video containing the activity.

APPENDIX D: RATING PROTOCOL

RATING PROTOCOL STEPS

1. Open the Excel Spreadsheet entitled RATING SHEETS_XX (your initials), and check the bottom tabs to toggle between the two test versions.

2. Read the directions at the top of the document.

3. Open the audio file that corresponds to the 8-digit number (starting with 6XXXXXXX) listed above each column.

4. Fast forward to approximately the 4-minute mark of the audio track to find the first word to be rated.

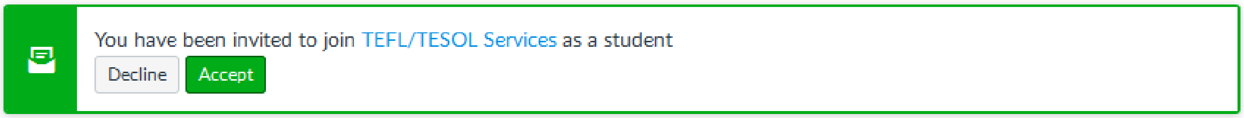
5. Listen carefully to the entire recording to rate each of the 36 words. Rate the pronunciation of the INITIAL PHONEME of each word (underlined on the rating sheet). The pronunciation should approximate Standard American English (SAE).
    - If the sound approximates the SAE sound, type the number “1” in the box next to the word (**Column Y/N**).
    - If the sound does not approximate the SAE sound, type the number “0” in the box next to the word.
    - Listen carefully for appropriate voicing and aspiration of initial sounds.
    - If the speaker makes a word plural, **rate as is**.
    - If the speaker says a word more than one time, **only rate the first utterance**.
    - If the speaker does not utter a word, type “N/A” in the box next to the word. **Do not assign a score to this word**.
    - If you cannot reliably determine the accuracy of a word (inaudible), type "N/A" in the box next to the word. **Do not assign a score to this word**.
    - If the participant says a completely different word, type that word in the box next to the word. **Do not assign a score to this word**.

6. Listen carefully to the entire recording to rate each of the 36 words.

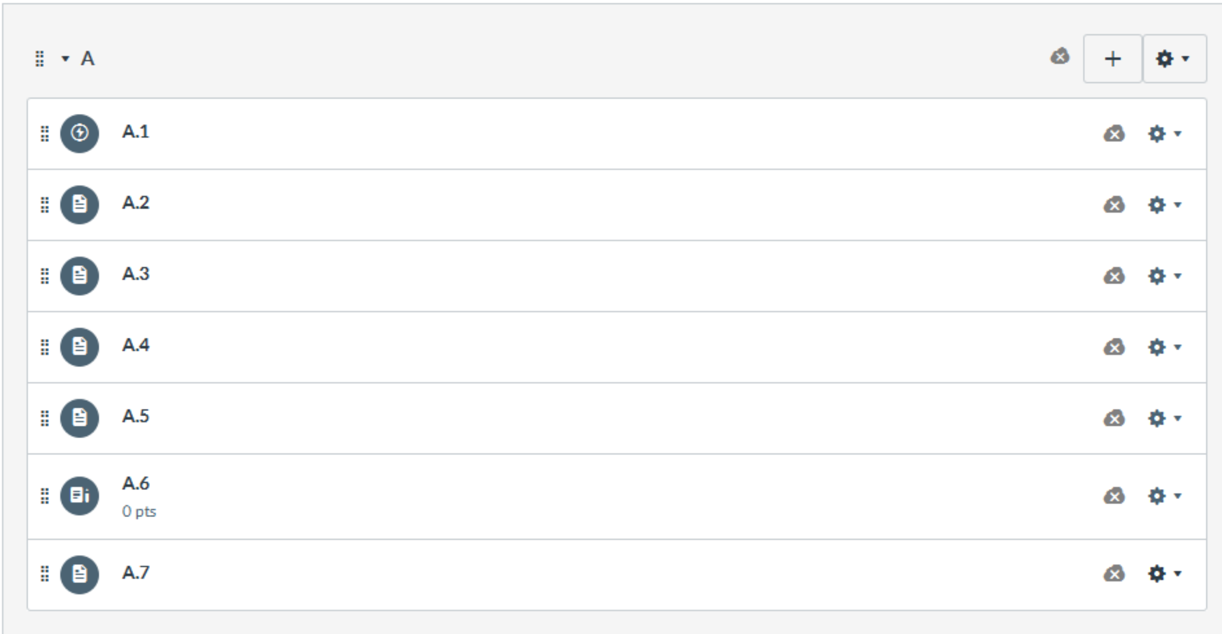
7. Repeat steps 3-6 for each audio recording. Remember to rate all of the audio files associated with each of the test versions (A and B). Save your work frequently.

APPENDIX E: DATA COLLECTION SCREENSHOTS

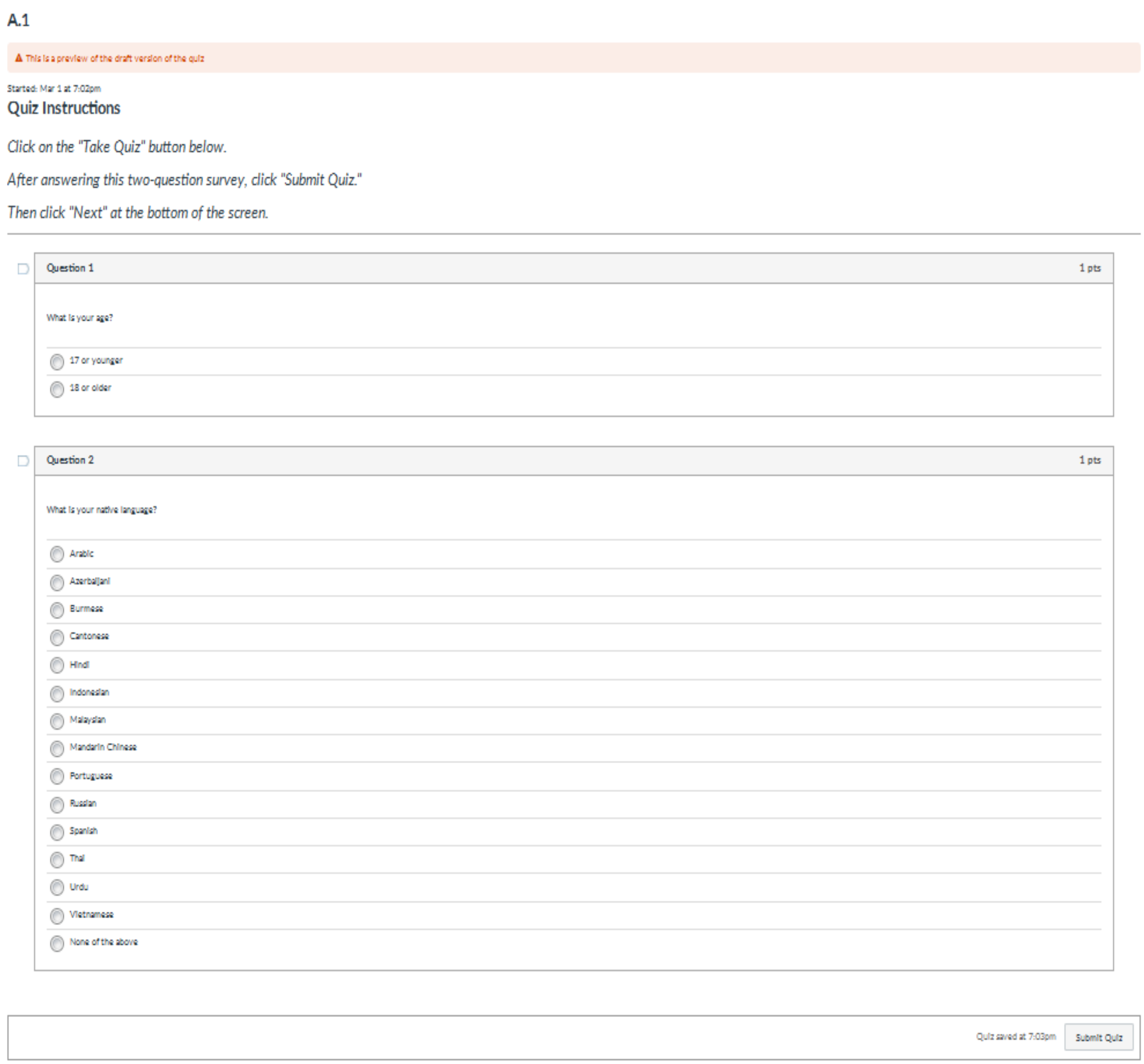
Consent Process



Module Overview



A.1: Demographic Questionnaire



A.2: Content Page

A.3: Content Page

A.3

Before beginning, please watch the short tutorial on the "Next" screen.

◂ Previous   Next ▸

A.4: Tutorial Video

A.4

After watching the tutorial, click "Next."

◂ Previous   Next ▸

A.5: SEMi, Version A

A.5

After completing the activity within this video, click "Next."

◂ Previous   Next ▸

A.6: Content Page

A.6

Stop your audio recorder now (Do NOT rename or delete any files).

*Close the **Voice Recorder** program NOW.*

Click "Next."

◂ Previous   Next ▸

A.7: Submit Assignment Page

A.7

Publish Edit

*Click "**Submit Assignment**."*

*Click "**Browse...**"*

*Choose "**Documents**" from the left side bar.*

*Click on the "**Sound recordings**" folder.*

*Select "**Recording (2)**" to attach your sound file. (The file should appear next to the "Browse..." button)*

Raise your hand if you have any questions.

*When ready, click "**Submit Assignment**."*

Then click "Next" to complete the activity.

Points 0

Submitting a file upload

Due	For	Available from	Until

A.8: Content Page

A.8

This concludes the activity. Please log off of Webcourses NOW.

◂ Previous Next ▸

APPENDIF F: SEMI, VERSION A

SPEECH ELICITATION METHOD INSTRUMENT (SEMI)

Standard American English (SAE) Version

(NOSE)

SPEAK EACH WORD ONE TIME AFTER YOU SEE OR HEAR IT

RING

NAIL

(MOON)

(PEN)

CAT

COW

(BALL)

(DRUM)

HAND

HORSE

(DUCK)

(CAT)

SKIRT

SOCKS

(COW)

(HAND)

FOOT

FORK

(HORSE)

(SKIRT)

PEN

BALL

(SOCKS)

(FOOT)

DRUM

DUCK

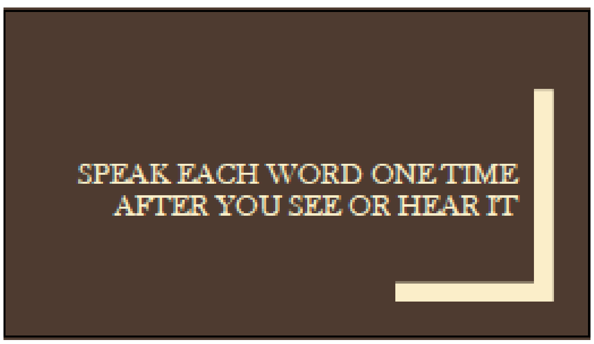

(FORK)

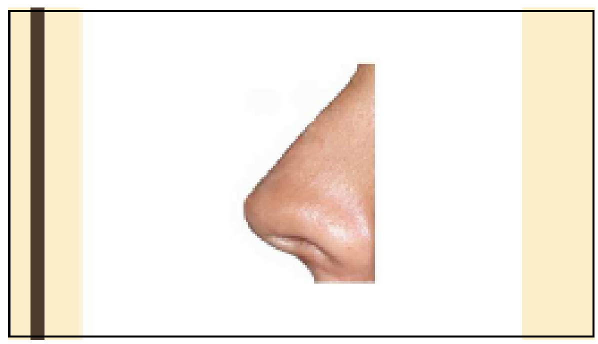

END OF ACTIVITY

APPENDIX G: SEMI, VERSION B

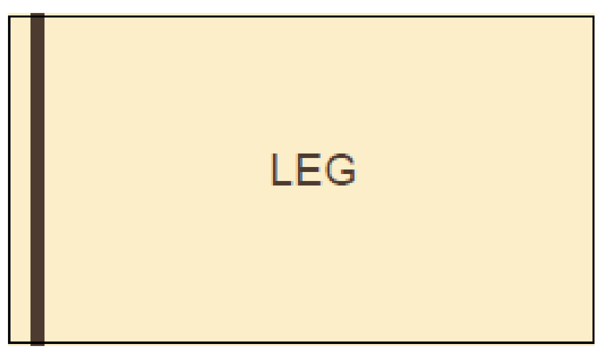

SPEECH ELICITATION METHOD INSTRUMENT (SEMI)
Standard American English (SAE) Version


(RING)


SPEAK EACH WORD ONE TIME AFTER YOU SEE OR HEAR IT




LAMP


LEG

(NAIL)

(SKIRT)

PEN

BALL

(SOCKS)

(FOOT)

DRUM

DUCK

(FORK)

(PEN)

CAT

COW

(BALL)

(DRUM)

HAND

HORSE

(DUCK)

(CAT)

SKIRT

SOCKS

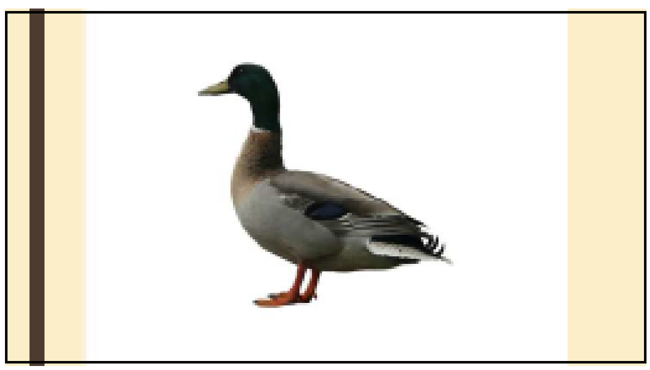
(COW)

(HAND)

FOOT

FORK


(HORSE)




END OF ACTIVITY

REFERENCES

Adlington, R. L., Laws, K. R., & Gale, T. M. (2009). The Hatfield Image Test (HIT): A new picture test and norms for experimental and clinical use. *Journal of Clinical & Experimental Neuropsychology, 31*(6), 731-753. doi:10.1080/13803390802488103

Akamatsu, N. (1999). The effects of first language orthographic features on word recognition processing in English as a second language. *Reading & Writing: An Interdisciplinary Journal, 11*(4), 381–403.

Asher, J. J., & Garcia, R. (1969). The optimal age to learn a foreign language. *Modern Language Journal 53*, 334–341. doi: 10.1111/j.1540-4781.1969.tb04603.x

Baddeley, A. D. (1986). *Working memory.* Oxford, England: Oxford University Press.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York: Academic Press.

Baetens-Beardsmore, H. (1979). The recognition and tolerance level of bilingual speech. *Working Papers on Bilingualism Toronto*, (19), 115-127.

Bailey, B. (2000). Language and negotiation of ethnic/racial identity among Dominican Americans. *Language in Society*, 29, 555–582.

Bankson, N. W., & Bernthal, J. E. (1990). Bankson-Bernthal Test of Phonology [Measurement instrument]. Austin, TX: Pro-Ed.

Bayard, D., & Green, J. A. (1995). Evaluating English accents worldwide. *Te Reo, 48,* 21-28.

Beebe, L. (1977). The influence of the listener on code-switching. *Language Learning*, *27*(2), 331-339. doi: 10.1111/j.1467-1770.1977.tb00125.x

Beebe, L. (1980). Sociolinguistic variation and style shifting in second language acquisition. *Language Learning, 30*(2), 433-447.

Biber, D., & Conrad, S. (2001). Quantitative corpus-based research: Much more than bean counting. *TESOL Quarterly, 35*, 331-336. doi: 10.2307/3587653

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow, UK: Longman.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511804489

Biber, D., & Reppen, R. (2002). What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition, 24*, 199-208. doi: 10.1017/S0272263102002048

Birdsong, D. (Ed.). (1999). *Second language acquisition and the critical period hypothesis.* Mahwah, NJ: Lawrence Erlbaum Associates.

Block, D. (2006). *Multilingual identities in a global city: London stories.* London: Palgrave. doi: 10.1057/9780230501393

Bley-Vroman, R. (1989). What is the logical problem of foreign language learning? In S. Gass & J. Schachter (Eds.), *Linguistic perspectives on second language acquisition* (pp. 41-68). Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139524544

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E. Tarone, S. Gass, & A. Cohen. (Eds.), *Research methodology in second-language acquisition* (pp. 245-261). Hillsdale, NJ: Lawrence Erlbaum.

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition, 19*(4), 447-465. doi: 10.1017/S0272263197004026

Borden, G., Gerber, A., & Milsark, G. (1983). Production and perception of the /r/-/l/ contrast in Korean adults learning English. *Language Learning*, *33*(4), 499-526. doi:10.1111/j.1467-1770.1983.tb00946.x

Bornstein, M. H. (Ed.) (1987). *Sensitive periods in development: Interdisciplinary perspectives.* Hillsdale, NJ: Lawrence Erlbaum.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*(4), 2299-2310. doi: 10.1121/1.418276

Brennan, E. M., & Brennan, J. S. (1981). Accent scaling and language attitudes: Reactions to Mexican American English speech. *Language & Speech*, *24*(3), 207-221.

Bresnahan, M. J., Ohashi, R., Nebashi, R., Liu, W. Y., & Shearman, S. M. (2002). Attitudinal and affective response toward accented English. *Language & Communication*, *22*, 171-185. doi: 10.1016/S0271-5309(01)00025-8

Breitkreutz, J., Derwing, T. M., & Rossiter, M. J. (2002). Pronunciation teaching practices in Canada. *TESL Canada Journal*, *19*, 51-61. doi: 10.18806/tesl.v19i1.919

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS ONE, 5*(5), 1-13. doi: 10.1371/journal.pone.0010773

Brysbaert, M., & New, B. (2009). Moving beyond Kučera 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. doi: 10.3758/BRM.41.4.977

Buchwald, A. B.,Winters, S. J., Pisoni, D. B. (2009).Visual speech primes open-set recognition of spoken words. *Language & Cognition Processes, 24*(4), 580–610. doi: 10.1080/01690960802536357

Catford, J. C. (1987). Phonetics and the teaching of pronunciation: A systematic description of English phonology. In J. Morley (Ed.), *Current perspectives on pronunciation: Practices anchored in theory* (pp. 83-100). Washington, DC: TESOL.

Celce-Murcia, M., Brinton, D., & Goodwin, J. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. New York: Cambridge University Press.

Chambers, J. K., & Trudgill, P. (1980). *Dialectology*. Cambridge: Cambridge University Press.

Chiba, R., Matsuura, H., & Yamamoto, A. (1995). Japanese attitudes toward English accents. *World Englishes, 14*(1), 77-86. doi: 10.1111/j.1467-971X.1995.tb00341.x

Chong, R. H., & Tan, Y. (2013). Attitudes toward accents of Mandarin in Singapore. *Chinese Language & Discourse, 4*(1), 120-140. doi: 10.1075/cld.4.1.04cho

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100*(4), 589-608. doi: 10.1037/0033-295X.100.4.589

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*(1), 204-256. doi: 10.1037/0033-295X.108.1.204

Compton, A. (1983). Compton Phonological Assessment of Foreign Accent [Measurement instrument]. San Francisco, CA: Carousel House.

Costa, A., Albareda, B., & Santesteban, M. (2008). Assessing the presence of lexical competition across languages: Evidence from the Stroop task. *Bilingualism: Language & Cognition*, *11*(1), 121-131. doi: 10.1017/S1366728907003252

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238. doi: 10.2307/3587951

Crystal, D. (1997). *English as a global language*. Cambridge: Cambridge University Press.

Curtiss, S. (1977). *Genie: A psycholinguistic study of a modern day "wild child"*. New York: Academic Press.

Dalton-Puffer, C., Kaltenboeck, G., & Smit, U. (1997). Learner attitudes and L2 pronunciation in Austria. *World Englishes, 16*(1), 115-128. doi: 10.1111/1467-971X.00052

Davies, M. (2015). *The corpus of contemporary American English (1990-2015).* Retrieved online at http://corpus.byu.edu/coca/

Dawson, J., & Tattersall, P. (2001). Structured Photographic Articulation Test II featuring Dudsberry [Measurement instrument]. DeKalb, IL: Janelle Publishing.

Denes, P. B. (1963). On the statistics of spoken English. *Journal of the Acoustical Society of America, 35*(6), 892-904. doi:10.1121/1.1918622

Derwing, T. M., & Munro, M.J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly, 39*(3), 379-397. doi: 10.2307/3588486

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*(4), 476-490. doi: 10.1017/S026144480800551X

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*(3), 393-410. doi: 10.1111/0023-8333.00047

Derwing, T. M., & Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning, 13*(1), 1-17.

Dickerson, L. (1975). The learner's interlanguage as a system of variable rules. *TESOL Quarterly, 9*(4), 401-407. doi: 10.2307/3585624

Dickerson, L., & Dickerson, W. (1977). Interlanguage phonology: Current research and future directions. The notions of simplification, interlanguages, and pidgins, and their relation to second language pedagogy. In *Actes du 5eme Colloque de Linguistique Appliquee de Neuchatel.* Geneva: Droz, 18-30.

Dodd, B., Hua, Z., Crosbie, S., Holm, A., & Ozanne, A. (2006). Diagnostic Evaluation of Articulation and Phonology [Measurement instrument]. San Antonio, TX: Pearson.

Dowd, J. L. (1985). Phonological variation in L2 speech: The effects of emotional questions and field-dependence/field-independence on second language performance. *Dissertation Abstracts International*, *46*(1), 138A-139A.

Elliott, A. R. (1995). Foreign language phonology: Field independence, attitude, and the success of formal instruction in Spanish pronunciation. *Modern Language Journal, 79*, 530–542. doi: 10.1111/j.1540-4781.1995.tb05456.x

Elliott, A. R. (1997). On the teaching and acquisition of pronunciation within a communicative approach. *Hispania*, *80*(1), 95-108. doi: 10.2307/345983

Erdener, V. D., & Burnham, D. K. (2005). The role of audiovisual speech and orthographic information in nonnative speech production. *Language Learning, 55*(2), 191-228. doi: 10.1111/j.0023-8333.2005.00303.x

Fayer, J. M., & Krasinski, E. K. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*, 313–326. doi:10.1111/j.1467-1770.1987.tb00573.x

Firth, S. (1992). Pronunciation syllabus design: A question of focus. In P. Avery & S. Ehrlich (Eds.), *Teaching American English pronunciation* (pp. 173-183). Oxford: Oxford University Press.

Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 229–273). Timonium, MD: York Press.

Flege, J. E. (2012). The role of input in second language (L2) speech learning. In *Proceedings of the 6th International Conference on Native and Nonnative Accents of English*, Łódź (Poland), 6-8 December 2012.

Flege, J. E., Frieda, E. M., & Nozawa, T. (1997). Amount of native language (L1) use affects the pronunciation of an L2. *Journal of Phonetics, 25*, 169–186. doi: 10.1006/jpho.1996.0040

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory & Language, 41*, 78-104. doi: 10.1006/jmla.1999.2638

Folse, K. S. (2004). *Vocabulary myths: Applying second language research to classroom teaching*. Ann Arbor: University of Michigan Press. doi: 10.3998/mpub.23925

Frith, U., Wimmer, H., & Landerl, K. (1998). Differences in phonological recoding in German- and English-speaking children. *Scientific Studies of Reading, 2*(1), 31–54. doi: 10.1207/s1532799xssr0201_2

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305-327. doi: 10.1093/applin/amt015

Gatbonton, E. (1975). *Systematic variations in second language speech: A sociolinguistic study*. Unpublished doctoral dissertation, McGill University, Montreal, Canada.

Gatbonton, E. (1978). Patterned phonetic variability in second-language speech: A gradual diffusion model. *Canadian Modern Language Review, 34*, 335–347.

Gatbonton. E., & Trofimovich, P. (2008). The ethnic group affiliation and L2 proficiency link: Empirical evidence. *Language Awareness, 17*(3), 229-248.

Gatbonton, E., Trofimovich, P., & Magid, M. (2005). Learners' ethnic group affiliation and L2 pronunciation accuracy: A sociolinguistic investigation. *TESOL Quarterly, 39*(3), 489–511. doi: 10.2307/3588491

Giles, H. (1970). Evaluative reactions to accents. *Educational Review, 22*(3), 221-227. doi:10.1080/0013191700220301

Goldman, R., & Fristoe, M. (2015). Goldman-Fristoe Test of Articulation (Version 3) [Measurement instrument]. Pearson Education.

Goldstein, B., Fabiano, L., & Iglesias, A. (2004). Spontaneous and imitated productions in Spanish-speaking children with phonological disorders. *Language, Speech & Hearing Services in Schools, 35*(1), 5-15. doi: 10.1044/0061-1461(2004/002)

Goldstein, B., & Iglesias, A. (2001). The effect of dialect phonological analysis: Evidence from Spanish-speaking children. *American Journal of Speech-Language Pathology, 10*(4), 394-406. doi: 10.1044/1058-0360(2001/034)

Goldstein, B., & Iglesias, A. (2006). Contextual Probes of Articulation Competence-Spanish [Measurement instrument]. Greenville, SC: Super Duper Publications.

Goldstein, T. (1997). *Two languages at work: Bilingual life on the production floor.* New York: Mouton de Gruyter. doi: 10.1515/9783110815221

Grant, L. (2014). *Pronunciation myths: Applying second language research to classroom teaching*. Ann Arbor: University of Michigan Press. doi: 10.3998/mpub.4584330

Guiora, A. Z., Beit-Hallahmi, B, Brannon, R. C. L., Dull, C. Y., & Scovel, T. (1972). The effects of experimentally induced changes in ego states on pronunciation ability in a second language: An exploratory study. *Comprehensive Psychiatry, 13*(5), 421-428. doi:10.1016/0010-440X(72)90083-1

Hamada, M., & Koda, K. (2010). The role of phonological decoding in second language word meaning inference. *Applied Linguistics 31*(4), 513-531. doi: 10.1093/applin/amp061

Hanulová, J., Davidson, D. J., & Indefrey, P. (2011). Where does the delay in L2 picture naming come from? Psycholinguistic and neurocognitive evidence on second language word production. *Language & Cognitive Processes*, *26*(7), 902-934. doi: 10.1080/01690965.2010.509946

Herschensohn, J. (2007). *Language development and age.* Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511486487

Hockett, C. (1967). The quantification of functional load. *WORD*, *23*, 320–339. doi: 10.1080/00437956.1967.11435484

Hodson, B. W. (2004). Hodson Assessment of Phonological Patterns (Version 3) [Measurement instrument]. Austin, TX: Pro-Ed.

Hoshino, N., & Kroll, J. F. (2008). Cognate effects in picture naming: Does cross-language activation survive a change of script? *Cognition*, *106*(1), 501-511. doi: 10.1016/j.cognition.2007.02.001

Indefrey, P., & Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. *Cognition 92*(1–2), 101–144. doi:10.1016/j.cognition.2002.06.001

Ivanova, I., & Costa, A. (2008). Does bilingualism hamper lexical access in speech production? *Acta Psychologica*, *127*(2), 277-288. doi: 10.1016/j.actpsy.2007.06.003

Jakobson, R. (1931). Prinzipien der historischen phonologie. *Travaux du cercle linguistique de Prague, 4*, 246–267. doi: 10.1515/9783110845174-005

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics, 23*, 83–103. doi: 10.1093/applin/23.1.83

Jiang, X. (2016). The role of oral reading fluency in ESL reading comprehension among learners of different first language backgrounds. *Reading Matrix: An International Online Journal, 16*(2), 227-242.

Kang, O., & Moran, M. (2014). Functional loads of pronunciation features in nonnative speakers' oral assessment. *TESOL Quarterly, 48*(1), 176-187. doi: 10.1002/tesq.152

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *Modern Language Journal, 94*(4), 554-566. doi: 10.1111/j.1540-4781.2010.01091.x

Kanno, Y., & Norton, B. (2003). Imagined communities and educational possibilities: Introduction. *Journal of Language, Identity, & Education, 2*(4), 241-249. doi: 10.1207/S15327701JLIE0204_1

Kashru, B. B. (1997). World Englishes 2000: Resources for research and teaching. In L.E. Smith & M. L. Forman (Eds.), *World Englishes 2000* (pp. 209-251). Honolulu: University of Hawaii Press.

Khan, L., & Lewis, N. (2002). Khan-Lewis Phonological Analysis (Version 2) [Measurement instrument]. San Antonio, TX: Pearson.

Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition, 93*(1), B39-B47. doi: 10.1016/j.cognition.2003.11.003

King, R. D. (1967). Functional load and sound change. *Language, 43*(4), 831-852. doi: 10.2307/411969

Kirk, C., & Vigeland, L. (2014). A psychometric review of norm-referenced tests used to assess phonological error patterns. *Language, Speech, & Hearing Services in Schools, 45*(4), 365-377. doi: 10.1044/2014_LSHSS-13-0053

Koda, K. (2016). Development of word recognition in a second language. In X. Chen, V. Dronjic, R. Helms-Park, (Eds.). *Reading in a second language: Cognitive and psycholinguistic issues* (pp. 70-98). New York, NY: Routledge/Taylor & Francis Group.

Krashen, S. (1978). Individual variation in the use of the monitor. In W. Ritchie (Ed.), *Second Language Acquisition Research: Issues & Implications.* New York: Academic Press.

Krashen, S. (1982). *Principles and practice in second language acquisition*. New York: Prentice-Hall.

Kroll, J. F., Bobb, S. C., & Wodniecka, Z. (2006). Language selectivity is the exception, not the rule: Arguments against a fixed locus of language selection in bilingual speech. *Bilingualism, Language & Cognition, 9*(2), 119-135. doi: 10.1017/S1366728906002483

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Labov, W. (1966). *The social stratification of English in New York City*. Arlington, VA: Center for Applied Linguistics. doi: 10.1017/CBO9780511618208

Landerl, K. (2000). Influences of orthographic consistency and reading instruction on the development of nonword reading skills. *European Journal of Psychology of Education, 15*(3), 239–257. doi: 10.1007/BF03173177

Lenneberg, E. (1967). *Biological foundations of language.* New York: Wiley.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral & Brain Sciences*, *22*(1), 1–38.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369-377. doi: 10.2307/3588485

Lippi-Green, R. (1997). *English with an accent: Language, ideology, and discrimination in the United States* (2nd ed.). London, England: Routledge.

Lindemann, S. (2010). Who's "unintelligible"? The perceiver's role. *Issues in Applied Linguistics*, *18*(2), 223-232.

Liu, Q. (2011). Factors influencing pronunciation accuracy: L1 negative transfer, task variables and individual aptitude. *English Language Teaching*, *4*(4), 115-120. doi: 10.5539/elt.v4n4p115

Liu, Q., & Fu, Z. (2011). The combined effect of instruction and monitor in improving pronunciation of potential English teachers. *English Language Teaching*, *4*(3), 164-170. doi: 10.5539/elt.v4n3p164

MacDonald, S. (2002). Pronunciation – views and practices of reluctant teachers. *Prospect 17*(3), 3–18.

Major, R. C. (1987). A model for interlanguage phonology. In G. Ioup & S. H. Weinberger (Eds.), *Interlanguage phonology: The acquisition of a second language sound system* (pp. 101–125). New York: Newbury House/Harper & Row.

Major, R. C. (2002). The phonology of the L2 user. In V. Cook (Ed.), *Portraits of the L2 user* (pp. 67–92). Clevedon, UK: Multilingual Matters.

Mathesius, V. (1931). Zum problem der belastungs und kombinationsfahigkeit der phoneme. *Travaux du Cercle Linguistique de Prague*, *4*, 148–152.

Moreno-Martinez, F. J., & Montoro, P. R. (2012). An ecological alternative to Snodgrass & Vanderwart: 360 high quality colour images with norms for seven psycholinguistic variables. *PLoS ONE, 7*(5), 1-9.

Morrison, J.A., & Shriberg, L. D. (1992). Articulation testing versus conversation speech sampling. *Journal of Speech & Hearing Research, 35*, 259-273.

Morton, E. S., Brundage, S. B., & Hancock, A. B. (2010). Validity of the proficiency in oral English communication screening. *Contemporary Issues in Communication Science & Disorders*, *37*, 153-166.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165-178. doi: 10.1037/h0027366

Morton, J. (1980). The logogen model and orthographic structure. In U. Frith (Ed.). *Cognitive processes in spelling.* London: Academic Press.

Moyer, A. (1999). Ultimate attainment in L2 phonology: The critical factors of age, motivation, and instruction. *Studies in Second Language Acquisition 21*(1), 81–108. doi: 10.1017/S0272263199001035

Moyer, A. (2004). *Age, accent, and experience in second language acquisition*. Clevedon, England: Multilingual Matters.

Munro, M. J., & Derwing, T. M. (1994). Evaluations of foreign accent in extemporaneous and read material. *Language Testing*, *11*(3), 253-266. doi:10.1177/026553229401100302

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73-97. doi: 10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System, 34*(4), 520-531. doi: 10.1016/j.system.2006.09.004

Murphy, J. M. (1991). Oral communication in TESOL integrating speaking, listening, and pronunciation. *TESOL Quarterly*, *25*(1), 51-75. doi: 10.2307/3587028

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary.* Boston, MA: Heinle Cengage Learning.

Norton, B. (2000). *Identity and language learning.* London: Longman.

Norton, B. (2013). *Identity and language learning: Extending the conversation* (2nd ed.). Bristol, UK: Multilingual Matters.

Norton Pierce, B. (1995). Social identity, investment, and language learning. *TESOL Quarterly*, *29*, 9-31.

Olson, L. L., & Samuels, S. J. (1973). The relationship between age and accuracy of foreign language pronunciation. *Journal of Educational Research*, *66*(6), 263-268. doi: 10.1080/00220671.1973.10884476

Oyama, S. (1976). A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research*, *5*(3), 261-283. doi: 10.1007/BF01067377

Pavlenko, A., & Blackledge, A. (Eds.). (2004). *Negotiation of identities in multilingual settings.* Clevedon, UK: Multilingual Matters.

Pavlenko, A., & Norton, B (2007). Imagined communities, identity, and English language teaching. In J. Cummins & C. Davison (Eds.). *International handbook of English language teaching* (pp. 669-680). New York, NY: Springer.

Pennington, M. C., & Richards, J. C. (1986). Pronunciation revisited. *TESOL Quarterly, 20*(2), 207-25. doi: 10.2307/3586541

Pike, E. V. (1959). A test for predicting phonetic ability. *Language Learning*, *9*(1/2), 35-41. doi:10.1111/j.1467-1770.1959.tb01127.x

Piper, T., & Cansin, D. (1988). Factors influencing the foreign accent. *Canadian Modern Language Review, 44*, 334–42.

Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics, 29*(2), 191-215. doi: 10.1006/jpho.2001.0134

Polat, N., & Mahalingappa, L. J. (2010). Gender differences in identity and acculturation patterns and L2 accent attainment. *Journal of Language, Identity, & Education, 9*(1), 17-35. doi: 10.1080/15348450903476832

Purcell, E. T., & Suter, R. W. (1980). Predictors of pronunciation accuracy: A reexamination. *Language Learning,* 30(2), 272-287. doi: 10.1111/j.1467-1770.1980.tb00319.x

Pye, C., Ingram, D., & List, H. (1987). A comparison of initial consonant acquisition in English and Quiché. In K. Nelson, & A. Van Kleek (Eds.). *Children's Language, 6* (pp. 175-190). Hillsdale, NJ: Lawrence Erlbaum.

Qu, Q., Zhang, Q., & Damian, M. F. (2016). Tracking the time course of lexical access in orthographic production: An event-related potential study of word frequency effect in written picture naming. *Brain & Language, 159*, 118-126. doi: 10.1016/j.bandl.2016.06.008

Rajadurai, J. (2007). Intelligibility studies: A consideration of empirical and ideological issues. *World Englishes*, *26*(1), 87-98. doi: 10.1111/j.1467-971X.2007.00490.x

Reppen, R., Fitzmaurice, S. M., & Biber, D. (2002). *Using corpora to explore linguistic variation*. Amsterdam/Philadelphia: John Benjamins Publishing. doi: 10.1075/scl.9

Riney, T. J., & Flege, J. E. (1998). Changes over time in global foreign accent and liquid identifiability and accuracy. *Studies in Second Language Acquisition, 20*(2), 213-243. doi: 10.1017/S0272263198002058

Rosenthal, J., & Ehri, L. C. (2008). The mnemonic value of orthography for vocabulary learning. *Journal of Educational Psychology, 100*(1), 175-191. doi: 10.1037/0022-0663.100.1.175

Scales, J., Wennerstrom, A., Richard, D., & Wu, S. H. (2006). Language learners' perceptions of accent. *TESOL Quarterly, 40*(4), 715-738. doi: 10.2307/40264305

Schmidt, A. M., & Sullivan, S. (2003). Clinical training in foreign accent modification: A national survey. *Contemporary Issues in Communication Science & Disorders, 30*, 127-135.

Schmitt, B. M., Münte, T. F., & Kutas, M. (2000). Electrophysiological estimates of the time course of semantic and phonological encoding during implicit picture naming. *Psychophysiology 37*(4), 473-484. doi: 10.1111/1469-8986.3740473

Schumann, J. H. (1975). Affective factors and the problem of age in second language acquisition. *Language Learning, 25*(2), 209-235. doi: 10.1111/j.1467-1770.1975.tb00242.x

Schumann, J. H. (1986). Research on the acculturation model for second language acquisition. *Journal of Multilingual & Multicultural Development*, *7*(5), 379-392. doi: 10.1080/01434632.1986.9994254

Scovel, T. (1988). *A time to speak: A psycholinguistic inquiry into the critical period for human speech.* Cambridge, MA: Newbury House.

Secord, W., & Donahue, J. (2014). Clinical Assessment of Articulation and Phonology (Version 2) [Measurement instrument]. Greenville, SC: Super Duper Publications.

Secord, W. A., & Shrine, R. E. (1997). Secord Contextual Articulation Tests [Measurement instrument]. Greenville, SC: Super Duper Publications.

Service, E., Yli-Kaitala, H., Maury, S., & Kim, J. (2014). Adults' and 8-year-olds' learning in a foreign word repetition task: Similar and different. *Language Learning*, *64*(2), 215-246. doi: 10.1111/lang.12051

Shadish, W. R., Cook, T.D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for general causal inferences.* Belmont: CA: Wadsworth Cengage Learning.

Sholl, A., Sankaranarayanan, A., & Kroll, J. F. (1995). Transfer between picture naming and translation: A test of asymmetries in bilingual memory. *Psychological Science, 6*(1), 45-49. doi: 10.1111/j.1467-9280.1995.tb00303.x

Sikorski, L. (1991). Proficiency in Oral English Communication Manual [Measurement instrument]. Santa Ana, CA: LDS & Associates.

Simon, E., Chambless, D., & Alves, U. K. (2010). Understanding the role of orthography in the acquisition of a non-native vowel contrast. *Language Sciences, 32*, 380-394. doi: 10.1016/j.langsci.2009.07.001

Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). *The Michigan corpus of academic spoken English*. Ann Arbor, MI: Regents of the University of Michigan.

Smith, M. C., & Magee, L. E. (1980). Tracing the time course of picture-word processing. *Journal of Experimental Psychology: General*, *109*, 373-392. doi: 10.1037/0096-3445.109.4.373

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory, 6*(2), 174-215. doi: 10.1037/0278-7393.6.2.174

Snow, C. E., & Hoefnagel-Höhle, M. (1977). Age differences in the pronunciation of foreign sounds. *Language & Speech*, *20*(4), 357-365.

Stroop, J. R. (1935). Studies of interference in series of verbal reactions. *Journal of Experimental Psychology*, *18*, 643-662.

Stuart, D. M., & Carrasco, M. (1993). Semantic component of a cross-modal Stroop-like task. *American Journal of Psychology*, *106*(3), 383-405. doi: 10.2307/1423183

Suter, R. W. (1976). Predictors of pronunciation accuracy in second language learning. *Language Learning*, *26*, 233-253.

Swan, M., & Smith, B. (2001). *Learner English: A teacher's guide to interference and other problems*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511667121

Teutsch-Dwyer, M. (2001). (Re)constructing masculinity in a new linguistic reality. In A. Pavlenko, A. Blackledge, I. Piller, & M. Teutsch-Dwyer (Eds.), *Multilingualism, Second*

Language Acquisition, & Gender (pp. 175-198). New York: Mouton de Gruyter. doi: 10.1515/9783110889406

Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning, 41*, 177–204.

Timmis, I. (2002). Native-speaker norms and international English: A classroom view. *ELT Journal 56*(3), 240-249. doi: 10.1093/elt/56.3.240

Trofimovich, P., Gatbontin, E., & Segalowitz, N. (2007). A dynamic look at L2 phonological learning: Seeking processing explanations for implicational phenomena. *Studies in Second Language Acquisition, 29*(3), 407-448. doi: 10.1017/S027226310707026X

van Els, T., & de Bot, K. (1987). The role of intonation in foreign accent. *Modern Language Journal, 71*(2), 147-155. doi: 10.2307/327199

Van Severen, L., Gillis, J., Molemans, I., Van den Berg, R., De Maeyer, S., & Gillis, S. (2013). The relation between order of acquisition, segmental frequency and function: the case of word-initial consonants in Dutch. *Journal of Child Language*, *40*(4), 703-740. doi: 10.1017/S0305000912000219

Viggiano, M. P., Vannucci, M., & Righi, S. (2004). A new standardized set of ecological pictures for experimental and clinical research on visual object processing. *Cortex*, *40*(3), 491-509. doi: 10.1016/S0010-9452(08)70142-4

Wang, M., & Koda, K. (2007). Commonalities and differences in word identification skills among learners of English as a second language. *Language Learning, 57*(1), 201-222.

Wassink, A. B., Wright, R. A., & Franklin, A. D. (2007). Intraspeaker variability in vowel production: An investigation of motherese, hyperspeech, and Lombard speech in Jamaican speakers. *Journal of Phonetics, 35,* 363-379. doi:10.1016/j.wocn.2006.07.002

West, M. (1953). *A general service list of English words.* London: Longman.

Wolk, L., & Meisler, A. W. (1998). Phonological assessment: A systematic comparison of conversation and picture naming. *Journal of Communication Disorders*, *31*(4), 291-313. doi: 10.1016/S0021-9924(97)00092-0

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning & Communication, 3*(2), 215-229.

Yook, C., & Lindemann, S. (2013). The role of speaker identification in Korean university students' attitudes toward five varieties of English. *Journal of Multilingual & Multicultural Development, 34*(3), 279-296. doi: 10.1080/01434632.2012.734509