THE EFEECTS OF RATER-SPECIFIC CHARACTERISTICS ON THE RATING OF FOREIGN ACCENT

by

Sahar Almohareb

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Linguistics

Committee:

____ Director

____

____

____ Department Chairperson

____ Program Director

____ Dean, College of Humanities
and Social Sciences

Date: ___________________________________ Summer Semester 2020
George Mason University
Fairfax, VA

The Effects of Rater-Specific Characteristics on the Rating of Foreign Accent

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Sahar Almohareb
Master of Arts
Florida International University, 2014
Bachelor of Arts
Al-Imam Mohammad Ibn Saud Islamic University, 2007

Director: Steven H. Weinberger
Department of Linguistics

Summer Semester 2020
George Mason University
Fairfax, VA

Dedication

Hussain, Meshal, and Yara!

Acknowledgements

I would like to express my sincere gratitude to my advisor, Dr. Steven Weinberger, for his enthusiasm for the project and for his support and encouragement. I also would like to thank my committee members Dr. Harim Kwon and Dr. Douglas Wulf for providing guidance and feedback throughout this project. I am thankful to Dr. Yuting Guo for her help with recruitment, and to all the participants in this dissertation. Finally, I wish to thank my loving and supportive husband, Hussain, and my two wonderful children, Meshal and Yara, who provide me with unending inspiration and unconditional love.

Table of Content

List of Tables

Page

List of Figures

Abstract

THE EFFECTS OF RATER-SPECIFIC CHARACTERISTICS ON THE RATING OF FOREIGN ACCENT

Sahar Almohareb, Ph.D.

George Mason University, 2020

Dissertation Director: Dr. Steven H. Weinberger

This dissertation examines how rater-specific characteristics affect the rating of foreign accent in English. Previous studies have focused on the effects of speaker-specific characteristics on the rating of foreign accents (e.g., Munro & Derwing, 1995; Flege et al., 1999; Yeni-Komshian et al., 2000; Piske et al., 2001). However, recent studies (e.g., Kang, 2012; Schoonmaker-Gates, 2012; Weber & Pollman, 2010; Hayes-Harb et al., 2008; Huang & Jun 2015) have shown that rater-specific characteristics may also influence the rating of accented speech, but their precise influences are not well understood. Specifically, there are conflicting results regarding the effects of linguistic training of the raters (trained vs. untrained), the raters' nativeness status (native vs. nonnative), and the nonnative rater-speaker shared L1 status (shared L1 vs. not shared L1) on the rating of foreign accent. Additionally, the relationship between the nonnative raters' degree of accentedness and their ratings of other's accented speech has not been previously explored.

Accordingly, this dissertation's primary focus is to investigate how these raters' factors affect their accent rating. In particular, it examines whether trained and un-trained native raters differ in their rating behaviors, whether native and nonnative raters rate accented speech differently, whether nonnative raters rate speech from speakers sharing the same L1 background with them differently from speakers from a different L1 background, and whether nonnative raters are influenced in their foreign accent ratings as a result of their degree of foreign accent, self-reported L1 use, and length of residence (LoR) in an English-speaking country.

In an online foreign accent rating experiment, trained and un-trained native English raters and naïve nonnative raters from Arabic and Mandarin L1 backgrounds rated the degree of foreign accentedness of 150 short English phrases extracted from the Speech Accent Archive (Weinberger, 2019) from native speakers and nonnative speakers from Arabic and Mandarin L1 backgrounds. Nonnative raters were also recorded reading a sample text, called the "Stella passage," to assess their foreign accent.

The results show that native English raters with linguistic training did not differ from un-trained native raters in their ratings of the degree of foreign accent and that their ratings were strongly correlated. As for the difference between native and nonnative raters, the results show that they differed in their rating behaviors. Overall, native raters always assigned lower accent ratings to all the speech samples than nonnative raters. However, in looking at the rater-speaker shared L1 status, an interesting pattern emerges. The results show that nonnative raters rated nonnative speech samples from different L1 backgrounds to have more foreign accent. In particular, while native raters and Arabic raters in the study rated Arabic-accented samples to have similar degrees of foreign accent, Mandarin raters rated Arabic-accented samples significantly differently from English and Arabic raters.

Similarly, while native raters and Mandarin raters rated Mandarin-accented samples to have comparable degrees of foreign accent, Arabic raters rated the Mandarin-accented samples significantly differently from English and Mandarin raters. Additionally, the results show that nonnative raters' degrees of foreign accent and self-reported L1 use did not significantly influence their ratings. However, raters' LoR significantly affected their ratings, where raters with shorter LoR generally assigned lower accent rating scores than raters with longer LoR.

These findings suggest that rater-specific characteristics, such as nativeness status and rater-speaker shared L1 status, can influence the rating of foreign accent, pointing to the complex nature of accent rating and the factors affecting it. This dissertation provides important implications for research on foreign accent rating, spoken language assessment, and language pedagogy. It also offers a foundation for further investigation of the rater-specific characteristics influencing foreign accent assessment, contributing to our understanding of speech perception more generally.

Chapter 1

Introduction

This dissertation examines the influence of rater-specific characteristics on the rating of foreign-accented speech. In particular, the aim is to examine how native and nonnative raters from different L1 backgrounds rate the degree of foreign accent in English speech produced by native and nonnative English speakers from different L1 backgrounds. It investigates the relationship between the accent ratings and rater-specific characteristics related primarily to the English status of the rater (native vs. nonnative), the native raters' linguistic training (trained vs. un-trained), and speaker-rater shared L1 status (shared L1 vs. not-shared L1). It also explores the possible effects of other factors such as the nonnative raters' degree of foreign accent, self-reported L1 use, and length of residence in an English-speaking country.

Spoken language contains a multitude of information about the speaker. In addition to the linguistic information (e.g., speech errors), listeners may use indexical information (i.e., non-linguistic information about the speaker's identity, such as age and gender) in the acoustic signal to make judgments about that speaker almost instantly. Such determinations and judgments include the speaker's gender, age, race, and emotions, among other things (see Borrie, McAuliffe, Liss, O'Beirne, & Anderson, 2013; Murray, & Arnott, 1993; Munson, McDonald, DeBoe, & White, 2006). One of the judgments listeners make concerns the speaker's language background, namely, whether that speaker is a native or nonnative speaker of the target language and the degree of foreign accent in the speech (Munro, 2018). However, many factors can affect these judgments, including characteristics specific to the speaker or the listener, resulting in a positive or negative evaluation of the speech (Munro, 2003; 2008).

While it is true that everyone has an accent, nonnative or "foreign" accent is an identifiable accent phenomenon in the sense that it is affected more by speaker-specific features (e.g., age of learning the L2, and length of residence in the L2 country) than those affecting native accents. The speaker-specific features can potentially lead to more biases towards the nonnative speakers and, in turn, it may affect social interactions, than it does to native speakers with regional accents. It is important to note that this dissertation does not argue that nonnative speakers should correct their accents; foreign accent is part of the nonnative speaker's identity and should be embraced. Acquiring a language as an adult comes with the consequence of "sounding different" from speakers who acquired the target language from birth. The phenomenon of "sounding-different" from native speakers of the target language is perceptual (Munro & Derwing 1995a; Derwing & Munro, 1997; Munro, 2008; 2011), and it is known in the literature as "foreign accent." A formal definition of foreign accent is introduced in Chapter 2.

A great deal of previous research has focused on speaker-specific characteristics, such as the age of onset and L1 use, and their effects on their degree of foreign accentedness (e.g., Major, 1987; Munro, 1993; Guion, Flege & Loftin, 2000; Munro & Derwing, 2001 among others) as well as the effects on phonetic characteristics on the degree of foreign accent, such as VOT, vowel qualities, and speech rate (Major, 1987; Riney & Takagi, 1999; Munro, 1993; Flege, Bohn & Jang, 1997; Munro & Derwing, 2001; Gao, 2019). Recently, however, research has started to focus on rater-specific characteristics, such as the L1 background of the rater, the rater's familiarity with accented speech, the nonnative raters' proficiency level in the L2, and how these characteristics may affect the raters' judgments of accented speech (e.g., MacKay et al., 2006; Jiang & Song, 2010; Jiao, Watson, Wong, Gnevsheva & Nixon, 2019).

This dissertation focuses on the relationship between rater-specific characteristics and the rating of foreign-accented speech. Previous research on this topic has provided significant insights, as will be seen in the next chapter. However, it is not without limitations. Accordingly, this dissertation builds upon previous findings and aims to answer some open questions and investigate new ones. For example, previous literature on the effect of the rater's L1 background on the rating of foreign-accented speech has yielded mixed findings as to whether there is an effect of the rater's L1 background when rating foreign-accented speech. The answer to this question has important implications for perception studies in second language acquisition because it would tell us about possible listeners' biases and their effects on the data.

Additionally, this issue is closely related to whether native and nonnative raters differ in their accented speech rating. Foreign accent studies have typically used native speakers as their raters or "judges," and the findings from the few studies that have included nonnative raters cannot be generalized. These two questions also have implications for spoken language assessment tests and whether the background of the raters (examiners in this case) could potentially affect their assessment of the test taker's speech.

Further, this dissertation examines a new question about the relationship between the nonnative raters' degree of foreign accent and their ratings of foreign-accented speech. This question is particularly important because it tells us whether the rater's judgments are independent of their accent. If the nonnative raters' ratings of accent speech are influenced by the raters' degree of foreign accent, then it begs the question of whether their judgments are considered an accurate representation of the degree of foreign accentedness of the rated speech.

Accordingly, by examining whether rater-specific characteristics affect nonnative accented speech judgments, this dissertation investigates how the L1 background of nonnative raters, who either share or do not share the same L1 language background as the speaker, influences their rating of foreign-accented English speech. It also examines whether native and nonnative raters rate foreign accent differently and whether native raters rate accented speech differently due to their (linguistic) training status as being trained raters or un-trained raters. Lastly, it explores whether the nonnative raters' degree of accentedness correlates with their accented speech ratings.

Examining how raters from different L1 backgrounds judge foreign-accented speech could have important implications for linguistic and spoken language assessment research. For example, this would contribute to our understanding of the differences between native and nonnative raters and the relationship between speech production and perception as it relates to second language acquisition in general. Further, examining the relationship between the raters’ language background and their ratings of accented speech could shed light on potential rater biases in high-stake speech assessment tests like the TOEFL and IELTS tests and, consequently, have important implications for how to address them. Ultimately, any speech assessment test aims to evaluate the speaker’s fluency, proficiency, and ability to communicate in the target language successfully. These tests are administered by examiners (typically one) following a standardized criterion. The examiners could be native or nonnative speakers of the target language. However, if there are potential biases toward or against a specific accent due to the examiner’s language background, these may favor or harm the test taker’s assessment. In turn, this speaks to the bigger question of who the optimal judge is to assess speaking skills.

The rest of this dissertation is structured as follows: Chapter 2 provides a discussion of the relevant concepts and reviews the previous literature. Chapter 3 constitutes the methodology section for the rating experiment. Chapter 4 reports on the results and findings of the study. Lastly, Chapter 5 provides a general discussion and major conclusions.

Chapter 2

Literature Review

2.1. The Notion of Perceived Native-ness

The notion of accentedness is closely related to the notion of perceived native-ness. For example, judging speech to have a regional accent (e.g., a Southern American English accent) presumes the speaker to be a native speaker of English. However, judging speech to have a nonnative foreign accent (e.g., English with a Russian accent) presumes that the speaker is a nonnative speaker of the target language. Therefore, one cannot formulate a reasonable definition of foreign accent without a good understanding of what makes a person a native speaker of the target language. This is particularly relevant because our perception of the notion of native-ness is generally based on our knowledge of the real world, such as our knowledge of geography and how people in a particular area speak. Although it may seemingly be an easy task, defining a native speaker proves to be challenging, and the importance of its implications are far from trivial (cf., Davies, 2004).

Despite this, the easiest way to conceptualize this issue is by identifying the features that intrinsically make an individual a native speaker of the target language. These features could be linguistic (e.g., phonetic features) or non-linguistic (e.g., age of acquisition and type of input). It is worth noting that although the degree of accentedness is judged on linguistic grounds, non-linguistic factors affect these judgments. That is to say, one is not judged to be nonnative simply because of his age of acquisition or the type of input received, but rather is judged based on the individual's speech on which these non-linguistic factors may have intrinsic effects. Subsequently, defining a nonnative speaker becomes relatively more straightforward since it could be understood by what it is not (Thomson, 2017). Such accent judgment is then inherently based on identifying differences between the speaker's speech and the listener's knowledge about the features defining native speech in the target language.

On the surface, one of the apparent differences between native and nonnative speakers is the fact that native speakers do not have foreign accents, and nonnative speakers do (Scovel, 1969). However, this is oversimplified and may not be enough to differentiate these two groups of speakers. A commonly used distinction is generally made based on the age at which a language is acquired and its consequence on the speech and how the judges perceive it. This, of course, assumes that language is acquired, and there is a sensitive period within which native competence can be attained.

It is often argued that a native speaker is an individual who acquired a language during early childhood, assumed to starts from birth, and who has maintained using the language (Scovel, 1969; Davies, 2004; McArthur, 1992). In this sense, a nonnative speaker could be defined as an individual who acquired a language after early childhood. However, this definition is vague and further complicated by whether there are different cut-off ages for attaining native competence in different language domains (e.g., phonology, grammar, etc.) in the target language (cf., Seliger, 1978). However, it is overwhelmingly evident that while learning a language as an adult comes with the consequence of a foreign accent, acquiring a language as a child typically does not (Scovel, 1969; Coulmas, 1981; Medgyes, 1992).

Although there is no agreed-upon consensus on the cut-off age for the native acquisition of phonology, different researchers argue that it is much earlier than other language domains, generally around six years of age (e.g., Yeni-Komshian et al., 2000). Experimental evidence from Yeni-Komshian, Flege, and Liu (2000) suggests that adult speakers who started learning the target language between the ages 1-5 years are likely to be rated as native speakers in terms of pronunciation and the older the age of learning, the heavier the perceived accent will become gradually evident. In addition to language acquisition during early childhood, Davies (1991, 2003, 2004) lists other features of a native speaker, including the intuitive knowledge about idiolectal and standard grammar, ability to produce effortless and fluent discourse, and the capacity to communicate competently.

Indeed, acquiring a language as a typically developing child with no hearing loss or genetic impairments has no limitation, but learning it as an adult comes with restrictions that are likely due to the human brain's nature, as Scovel (1969) notes. Although the construct of who should be regarded as a native speaker could theoretically mean different things from different perspectives (Davies, 2004), for our purposes, it seems that age of acquisition is the most critical feature distinguishing native from nonnative speakers in terms of pronunciation, all other things being equal. Accordingly, this dissertation considers a *native speaker* to be an individual who acquired the target language naturalistically from birth and continued to use it and is judged to be a native speaker by listeners. Conversely, in this dissertation, a nonnative speaker of the target language is defined as an individual who started learning the target language after early childhood, after acquiring a first language from birth and is judged to be a nonnative speaker by listeners. The final stipulation that the individual's accent is judged by listeners is included in the definitions because the phenomenon of accent is perceptual. For example, there indeed exist, individuals, who have acquired a second language after age 6, but who nonetheless have managed to achieve phonological production that is judged to be exceptionally native-like. Crucially, even though such individuals exist, they are in the distinct minority of individuals who started learning the target language after 6.

2.2. Accentedness in the L2 (Foreign Accent)

Defining the term *accentedness* is no less challenging than defining the notion of *nativeness*. One reason why defining *accentedness* can be challenging is that it is often conflated with other related, yet independent concepts, namely *comprehensibility*, and *intelligibility*. *Comprehensibility* is defined as "the listener's estimation of difficulty in understanding an utterance" (Munro et al., 2006: 112). *Intelligibility* is defined as "the extent to which a speaker's message is actually understood by a listener" (Munro & Derwing, 1995: 76). However, unlike earlier studies, this dissertation only examines the perceived degree of foreign accent and does not attend to comprehensibility and intelligibility. As shown in Chapter 3, comprehensibility and intelligibility are controlled by providing the raters with the scripts of what is spoken in the audio files. This obviates the need to test these two concepts.

Another challenge when defining the term *accentedness* relates to finding a precise and explicit definition that encompasses all aspects of the term. *Accentedness* is not restricted to nonnative speech since native speakers also have different regional accents (Munro, 2018). The following is a list of some of the previous definitions for the terms *foreign accent* and *global foreign accent*, followed by a discussion of their useful insights. These terms are often used interchangeably to refer to the same concept, in which case both terms refer to the overall evaluation of the degree of foreign accent in speech.

- Wayland (1997) defines foreign accent as "speech which differs acoustically from the native phonetic norm and is auditorily detectable by native speakers" (p. 346).
- Munro and Derwing (1995) define foreign accent as ''non-pathological speech that differs in some noticeable respects from native speaker pronunciation norms'' (p. 289).
- Munro (1998) defines it as "nonpathological speech produced by L2 learners that differs in partially systematic ways from the speech characteristics of native speakers." (p. 139).
- Derwing and Munro (2009) define it as "how different a pattern of speech sounds compared to the local variety" (p. 476).
- McCullough (2013) defines it as "the percept of deviations from a pronunciation norm that a listener attributes to the talker not speaking the target language natively" (p. 6).

- Isaacs and Thomson (2013) define it as “how different the speaker sounds from a NS” (p. 141).
- Jułkowska and Cebrian (2015) define it “as the listener’s perception of how closely the pronunciation of an L2 speaker mirrors the pronunciation of a native speaker of a given language” (p. 212).
- Saito, Trofimovich, and Isaacs (2017) define it as "listener’s perceptions of the degree to which L2 speech is influenced by his/her [the speaker’s] native language and/or colored by other nonnative features" (p. 8).
- Riney, Takada, and Ota (2000) define global foreign accent as “the degree to which an L2 speaker’s productions are perceived to differ from those of a native speaker” (p. 713).
- Major (2001) defines global foreign accent as “the overall impression concerning NSs form whether or not and to what degree a person sounds native or nonnative” (p. 19).

The previous definitions share the idea that foreign accent is related to pronunciation and that it is a perceptual phenomenon since it requires a judgment by the listener (also see Thomson, 2017). Most of these definitions refer to the native speakers’ pronunciation as the baseline for comparison, which reinforces that the notions of accentedness and nativeness are closely related. The exception to the previous statement is the definition proposed by Saito et al. (2016).

Here, it must be noted that although the notions of *accentedness* and *nativeness* of the speaker are related, they do not necessarily mean the same thing. McCullough (2013) argues that while *accentedness* is a judgment about the speech, *nativeness* is a judgment about the speaker. A person could be a native speaker of the target language, but his speech could reveal acoustic features specific to a regional accent. Further, a person could be a nonnative speaker of the target language, but his speech could reveal acoustic features that are foreign to the target language. Thus, a person can be judged native or nonnative because his speech can be judged to have regional or foreign characteristics, as the case may be. This judgment implies that the listener has learned geographic knowledge about accents. For example, to a native English speaker from Ohio, an Alabaman accent could be as distinctive as a Russian accent. Nevertheless, the Ohio listener will recognize the speaker with an Alabaman accent as a native English speaker and the Russian speaker speaking English as a nonnative English speaker.

In foreign accent research, foreign accentedness and non-nativeness are often conflated to refer to the rater's overall impression of whether a person is a native or nonnative speaker and how different a person sounds from native speakers (e.g., Major, 2001). Although *nativeness* is generally regarded as a binary notion, and *accentedness* is regarded as a continuum, a listener's judgment about one implies a judgment about the other. It is worth noting that the notion of a "heritage language speaker" might be argued as a possible middle ground between native and nonnative speakers. However, in this dissertation, the notion of nativeness is treated as binary. Specifically, nativeness is binary in the sense that an individual is either a native or nonnative speaker of the target language, and accentedness is a continuum in the sense that there are degrees of accentedness ranging from no accent to a heavy accent. Accordingly, a listener's judgment about the degree of accentedness, for example, implies a judgment about the nativeness status. In other words, gauging the accentedness necessarily implies a judgment about nativeness. Similarly, if a listener judged a speaker to be nonnative based on an auditory stimulus, it is because the listener perceived some degree of foreign accent in the speech.

In this dissertation, foreign accent is taken to be a perceptual phenomenon related to pronunciation in which the listeners detect productions that they impressionistically believe cannot be attributed to native-speaker norms but instead arising from the speaker's nonnative status. Additionally, the foreign accent's degree is determined by the listeners' overall impression of how different a production sounds compared to that of a native speaker. This definition assumes that an informed judgment about foreign accent requires the listener to be familiar with how native speakers of target language sound.

2.3. Measuring Foreign Accent

Foreign accent is highly variable and appears to be gradient due to the various factors influencing it. Therefore, any measurement method of the degree of foreign accent needs to reflect such variability. Accordingly, previous research has typically used rating scales or a continuum. The most common method employs a rating scale (often referred to as Likert-scale), where raters choose one number from a pre-determined range. Rating scales varied in terms of their points ranging from 0 to 1000 points, with scales with 5 and 9 points being the most common. Some scales have numbers for every scale point, while other scales only have the endpoints numbered.

Additionally, while some scales have smaller numbers signifying lesser foreign accent and higher numbers signifying more foreign accent, some scales have the opposite. Thomson (2017) notes that these differences in the implementation of these rating scales, as well as the lack of an explanation for picking a particular scale over the other, suggest that "there is no particular rationale for doing so" (p. 17). However, Southwood and Flege (1999) argue that using a 9-point (or 11-point) scale is favored because it may improve the rater’s sensitivity when scaling the degree of foreign accent. They argue that smaller scales may not have enough points for all raters to distinguish between accentedness degrees in different phrases. Table 1 lists some studies that measured the degree of foreign accent and how they implemented the scale; Table 1 is partly based on Thomson (2017).

Table 1
Sample foreign accent studies and their implemented scales:

Study	**Rating scale implemented**
Munro & Derwing (1995a)	1= no foreign accent, 9= very strong foreign accent
Southwood & Flege (1999)	1 = least accent, 7= most accent
Yeni-Komshian et al. (2000)	1= very strong accent, 9 = no accent
Piske, MacKay & Flege (2001)	1 = very strong foreign accent, 9 = no foreign accent
Kennedy & Trofimovich (2008)	1 = no nonnative accent, 9 = strong nonnative accent
Yuan, Jiang & Song (2010)	1 = no accent, 4, very strong accent
Trofimovich & Isaacs (2012)	1 = heavily accented, 9 = not accented at all
Isaacs & Thomson (2013)	1 = heavily accented, 5/9 = not accented at all
Kraut & Wulff (2013)	1= strong foreign accent, 7 = no foreign accent
O'Brien (2014)	1 = no accent, 9 = extremely strong accent
Jułkowska & Cebrian (2015)	1 = no foreign accent, 9 = strong foreign accent
Saito et al. (2015)	0–1000 with endpoints not reported
Huang & Jun (2015)	1 = strong foreign accent, 9 = native English speaker
Saito et al. (2016)	1 = no accent, 9 = heavily accented
Huang et al. (2016)	1 = strong foreign accent, 7= native English speaker

Following most previous literature, this dissertation uses a 9-point rating scale as a measurement method for the degree of foreign accent. As shown in Chapter 3, the rating scale is numbered from 1-9 with the endpoints labeled with 1 "no foreign accent" and 9 "strong foreign accent." The way the rating question is formulated motivated the decision to label 1 with "no foreign accent" and 9 with "strong foreign accent." The logic is that when rating the degree of foreign accent, the more foreign accent there is, the higher the number on the scale is to reflect it.

2.4. Factors Influencing the Rating of Foreign Accent

Many factors influence the rating of foreign accent, including phonetic characteristics in the speech signal, characteristics specific to the speaker, and characteristics specific to the rater. Although the focus of this dissertation is on the rater-specific characteristics and their effects of the rating of foreign accent, this section provides a short review of phonetic characteristics, followed by a brief overview of relevant studies that have focused on speaker-specific characteristics since these studies highlight important methodological considerations for the current study.

2.4.1. Phonetic Characteristics

In terms of the phonetic characteristics and their effects on the foreign accent, previous research has demonstrated that both segmental and suprasegmental characteristics affect the degree of foreign accent. For example, differences in the duration of the Voice Onset Time between native and nonnative speakers (e.g., Major, 1987; Riney & Takagi, 1999) as well as the differences in vowel qualities, such as vowel height and duration, between native and nonnative speakers (e.g., Munro, 1993; Flege, Bohn & Jang, 1997) contribute to the perception of foreign accent. Additionally, a slower speech rate was found to affect the rating of foreign accent (Munro & Derwing, 2001). More recently, Gao (2019) has conducted a comprehensive study on the accentedness rankings of various phonetic patterns in L2 speech.

2.4.2. Speaker-Specific Characteristics

Several rating studies that have examined the degree of foreign accent in relation to speaker-specific characteristics have demonstrated that specific characteristics contribute more to the perceived degree of foreign accent than others. These characteristics include the speaker's age of onset (i.e., the age at which one begins to acquire a language), age of arrival (AoA) (i.e., the age at which one arrives at a country where the target language is spoken), length of residence, type of L1, and L1 use (i.e., the self-reported frequency of native language use).

In their classic study, Munro and Derwing (1995) examined foreign accent, comprehensibility, and intelligibility. Their speech material was collected through a picture description task produced by two native English speakers and ten native Mandarin speakers, who were proficient L2 speakers of English. They were graduate students at a Canadian university with high TOEFL scores. The researchers evaluated the nonnative speakers' accents and determined that they ranged from moderately to heavily foreign-accented. This step was to ensure that there were various degrees of foreign accent in the data. For the foreign accent rating task, the researchers also opted to use native English speakers as their judges, employing a 9-points rating scale, where 1 = no foreign accent and 9 = very strong foreign accent. Accordingly, 18 native English-speaking raters, with basic knowledge of phonetics, rated 36 sentences. Raters completed a short practice session at the beginning of the session. The accent ratings showed that raters' ratings for accentedness were highly correlated, indicating high agreement and that their ratings were distributed across the different degrees of accentedness. Thus, the study found that raters were successful in identifying native and nonnative speakers and discerning the different degrees of foreign accent presented in the stimuli.

Flege, Yeni-Komshian, and Liu (1999) tested the global foreign accent rating in English produced by 240 Korean native speakers with varying ages of arrival (1-23 years) and 24 native English speakers. All the Korean participants were experienced English speakers with at least eight years of residence in the United States. The Korean participants were divided into 11 subgroups based on their arrival age with 2 to 3 years of increments between groups. In their study, the researchers recruited ten native English-speaking raters to rate 21 sentences on the degree of foreign accent using a 9-point rating scale. Raters completed a short practice session before the main experiment. Each sentence was judged three times by the raters. The results show that as the age of arrival increased, the degree of perceived degree of foreign accent increased. The researchers argued that the AoA effect on L2 phonology could be attributed to the sensitive period (i.e., whether the language is learned before a certain age) and perhaps L1 use for the Korean speakers.

In a follow-up study, Yeni-Komshian et al. (2000) also examined the effect of self-reported L1 use on the global foreign accent ratings by testing Korean-English bilinguals in both Korean (their L1) and English (their L2). They found that Korean-English bilinguals who reported using Korean more dominantly were rated by native English raters to have more foreign accent in their English speech than those who reported using English dominantly. Additionally, Korean-English bilinguals who used English dominantly were more likely to be rated by Korean native speakers to have some foreign accent in their Korean. This led Yeni-Komshian et al. (2000) to conclude that language dominance is another contributing factor to foreign accent perception.

Piske, MacKay, and Flege (2001) examined the perception of foreign accent based on the age of onset (i.e., age at which learning the L2 began) and L1 use in English by Italian immigrants to Canada. Their speakers were divided based on the age of onset (early and late bilinguals) and the amount of L1 (Italian) use (low-use vs. high-use). Speakers were asked to listen to question-answer stimuli and repeat the answer. In the rating task, native Canadian English raters were asked to rate the degree of perceived foreign accent on a 9-point rating scale, and the rating experiment included a short practice session. Results revealed that the speakers' age of onset was the most critical factor influencing the degree of global foreign accent. It was reported that late bilinguals were rated as having a stronger foreign accent than early bilinguals. Similarly, the amount of L1 use was found to influence the degree of foreign accent. Italian speakers who reported high L1 use were rated to have a more robust foreign accent in the L2 (English).

These studies, so far, share two methodological considerations. First, they used 9-point rating scales to collect foreign accent ratings. Methodologically, using rating scales seems to be the most suitable method for such data because of its efficiency, ease of use, and consistency. This is because accent judgment is a rapid, intuitive response to the stimulus, and the rating scale allows for such rapidity and efficiency. Rating scales also offer consistency of the measure used so that the scale can be easily quantifiable for statistical analysis. Second, these studies included a training session before the main experiment. This is done in order for raters to familiarize themselves with the task and the scale. For the present study, these two considerations are also employed. In this dissertation, a short training session for the raters was included so that participants became familiar with the rating process and were comfortable utilizing the different scale points. This study also used a 9-point rating scale to provide a sufficient range of scale points to accommodate the raters' choices.

However, a shortcoming of these studies was their focus on native raters. Relying only on native raters as the judges is incompatible, for example, with the practice of using nonnative examiners in high-stake speech assessment tests, such as the IELTS. For the IELTS pronunciation assessment, examiners are not necessarily native speakers of English. This is particularly common in test centers located in non-English speaking countries. Additionally, the idea that only native raters can serve as optimal judges of accentedness is not empirically motivated, especially in light of the diverse findings in this regard, as discussed in the subsequent section.

2.4.3. Rater-Specific Characteristics

In addition to the speaker-specific characteristics that have been reported in the literature, a growing body of research has focused on rater-specific characteristics that could influence the judgment of foreign-accented speech of nonnative speakers. These characteristics include the rater's experience with foreign-accented speech, and raters' L1 background and proficiency.

2.4.3.1. Native and Nonnative Raters

Most accent studies have generally focused on native raters as judges (see Derwing & Munro, 2015). This perhaps stems from the notion that native speakers are better than nonnative speakers at detecting the aspects of L2 speech that contribute to accent perception. In particular, native raters have full competence in the language. However, less is understood about nonnative raters and how they judge foreign accent compared with native raters. The studies that have included nonnative raters and compared their ratings to native raters have provided mixed findings (see Winke, 2013).

For example, some studies that have compared the rating differences between native and nonnative raters have shown that nonnative raters judged L2 speech more severely (i.e., assigned higher foreign accent scores) than native raters (e.g., Sheorey, 1985; Fayer & Krasinski, 1987; Caban, 2003; Kang, 2012; Schoonmaker-Gates, 2012). Other studies have found that nonnative raters judged L2 speech more leniently (i.e., assigned lower foreign accent scores) than native raters (Brown, 1995). In addition, other studies (Derwing & Munro, 2013; Flege, 1988; MacKay, Flege, & Imai, 2006; Kim, 2009; Zhang & Elder, 2011) have found no difference in the rating of L2 speech between native and nonnative raters. This confusion regarding native and nonnative raters requires further investigation.

Additionally, previous research has typically argued for an advantage of a shared L1 background. However, this factor has only been widely tested for intelligibility and comprehensibility, which leaves open whether this effect of shared L1 background also affects accentedness ratings. For example, Bent and Bradlow (2003) and Harding (2012) have reported a shared L1 advantage for nonnative raters with speakers from the same L1, as reflected in their intelligibility and comprehensibility ratings (see also Carey, Mannell, & Dunn, 2011; and Winke, Gass, & Myford, 2013 for similar findings). That is, raters found speech produced by speakers with whom they share an L1 more intelligible and easier to understand. In these studies, it is consistently found that nonnative raters tend to rate L2 speech from speakers sharing an L1 with them more leniently (i.e., more intelligible and more comprehensible) than other L2 speech. In summary, these studies demonstrate that, at best, the shared L1 background between the speaker and listener may facilitate a better understanding of the speech. It is worth noting that these two parameters (i.e., intelligibility and comprehensibility) are independent of accent rating, mainly because nonnative speech could be intelligible and comprehensible yet heavily accented.

In contrast, Munro, Derwing, and Morton (2006) asked raters from different L1 backgrounds (Cantonese, Japanese, Mandarin, and English) to rate accented English spoken by native Cantonese speakers, Japanese, Polish and Spanish. They found that the rating scores for intelligibility, comprehensibility, and accentedness from all the raters' groups were moderately to highly correlated, suggesting that raters did not generally benefit from a shared L1 background when rating intelligibility, comprehension, and accentedness. Similar findings have also been reported by Crowther, Isaacs, and Trofimovich (2016), where no differences were apparent between native and nonnative raters in their ratings of comprehensibility and accentedness. This lack of difference between native and nonnative raters with a shared L1 background with the speaker suggests that there may be other factors (such as proficiency) influencing the ratings. As Crowther et al. (2016) argue, this idea is consistent with the finding that the advantage of having a shared L1 background is limited to just low L2 proficiency (see also Hayes-Harb et al., 2008). In fact, in Munro et al. (2006) and Crowther et al. (2016), the nonnative raters were reported to have advanced or high intermediate proficiency in the target language (in this case English), which could potentially explain the lack of different ratings obtained from native and nonnative raters.

MacKay et al. (2006) tested nonnative English raters' ability to rate foreign accent in L2 English speech. The raters in this study had a different L1 background (in this case, L1 Arabic) from the nonnative speakers' (Italian). The results showed that ratings of nonnative raters were highly correlated with those of native English raters. This means that native and nonnative raters were consistently similar in their ratings, but this, of course, does not tell us whether they were identical. Nonetheless, this finding indicates that nonnative raters who do not share the same L1 background with the nonnative speakers can be as reliable as native raters. Reliability here refers to the fact that each rater's group, as a group, were consistent in the pattern of their ratings, and in turn, consistent with the rating pattern of the other rater's group.

Yuan, Jiang, and Song (2010) compared the foreign accent ratings in spontaneous English speech by English and Mandarin Chinese raters. The speech samples were selected from eight speakers from different L1 backgrounds (namely, Cantonese, Mandarin, Vietnamese, German, French, Spanish, Russian, and Japanese). The results showed that Mandarin raters were less sensitive to accented English speech than native

English raters. Mandarin raters' assessments of accented speech were lower (i.e., rated to have less foreign accent) than native English raters. Additionally, these Mandarin raters rated Mandarin and Cantonese accented speech as less accented than other speech samples from the other languages. This suggests that if the speaker and the rater share the same or similar L1 backgrounds (e.g., Mandarin raters rating Mandarin speakers or Mandarin raters rating Cantonese speakers), this might influence the rater's ratings of global foreign accent in the target language (L2).

In addition to the possible difference between native and nonnative raters and whether there is a shared L1 with nonnative raters, there is evidence from studies focusing on raters' feedback and interviews suggesting that native and nonnative raters may even differ in the strategies they use to reach their judgments of L2 speech, thus weighing features in the nonnative speech samples differently (Kim, 2009; Jun & Li, 2010; Zhang & Elder, 2011). For example, Jun and Li (2010) designed an accentedness and comprehensibility study to examine the factors that underlie native and nonnative raters' judgments of accented speech. The raters were asked to vocalize their thought processes when they assigned their ratings. They found that nonnative raters relied more on segmental and suprasegmental errors when rating the L2 speech. Native raters, by contrast, focused on more general considerations such as clarity of speech, ease of understanding, and the presence of specific speech errors (e.g., lisp) or stutter. They concluded that perhaps native speakers rely more on the overall impression of the speech instead of focusing on more fine-grained speech characteristics. In light of these findings, as Crowther, Trofimovich, and Isaacs (2016) argue, perhaps even if native and nonnative raters assign similar ratings to L2 speech, the factors underlying their assessment may be different.

To summarize, these studies offer mixed findings regarding the differences between native and nonnative raters, the shared L1 advantage between the speaker and the rater, and what strategies are used by native and nonnative raters when judging accented speech. This dissertation is concerned explicitly with the differences between native and nonnative raters and the effect of a shared L1 background between the rater and the speaker.

2.4.3.2. Rater’s Experience with Accented Speech and L2 Proficiency

In addition to the rater’s L1 background, some studies have examined the effect of the rater’s experience with accented speech on their ratings. Typically, experience with accented speech has been understood as the rater’s linguistic training, profession, or the familiarity with a specific speech community. The idea is that trained linguists, language teachers, and individuals who live in an immigrant community are likely to have more experience in detecting the degree of foreign-accented speech than judges lacking such experience. The nature of this experience is viewed as the extensive exposure to nonnative speech due to profession (e.g., ESL teachers), knowledge of foreign accents or exceptional sound systems (e.g., linguists), or immersion in a community where a specific nonnative accent is spoken (e.g., an Italian immigrant community).

For example, Flege et al. (1997) examined the ratings of two L1 English rater groups who judged L1 Italian learners of English. The raters differed in their level of experience with L2 Italian accented English speech. The experienced rater group consisted of monolingual native speakers of Canadian English who were familiar with Italian-accented English because of either living in an Italian immigrant community or working in a workplace with dominantly Italian immigrant workers. The inexperienced group included monolingual native speakers of Alabaman English who had no experience with this accent. The raters were asked to listen and rate the degree of foreign accent of English sentences produced by Italian immigrants to Canada with varying reported L1 use.

Although the results showed that the two native English rater groups were not very different in the way they rated the speakers, there was nonetheless an overall small difference between the two rater groups that was statistically significant. Both rater groups rated Italian-English speakers with low L1 use (Italian) to have less foreign accent in English, and those with high L1 use to have more foreign accent. However, experienced raters familiar with the Italian foreign accent were slightly better at detecting this foreign accent. That is, experienced raters were better than inexperienced raters at detecting the foreign accent even when the speakers started learning English as children and reported low use of their L1. The reason that the Canadian raters were better at detecting this was argued to be their familiarity with the range of accented speech, and, in this case, experience with the Italian accent. However, this better score might have been partly or entirely because the nonnative speakers were learning Canadian English.

Weber and Pollman (2010) also examined whether familiarity with foreign accent plays a role in English and Dutch speech ratings. They asked three groups of raters (L1 Dutch/L2 English, L1 German/ L2 English, who also knew Dutch, and L1 German/L2

English, who did not know Dutch) to rate the degree of foreign accent in 12 short English and Dutch sentences produced by six native and six nonnative speakers of English and Dutch. They used a 9-point rating scale. For the English sentences, German raters who did not have any familiarity with the Dutch accent performed equally as the Dutch raters and the German raters with familiarity with the Dutch accent. However, for the Dutch sentences, it was found that German raters who knew Dutch were better in detecting foreign accent than German raters who did not know Dutch. Nonetheless, German raters without Dutch familiarity were still able to detect nonnative Dutch speech.

This study suggests that raters' foreign accent ratings without any familiarity with the target language are comparable to those raters with familiarity with Dutch. This simply means that the degree of agreement among raters from both groups was high. In their study, familiarity with the target language (not the speaker's native language) increases the similarity of foreign accent ratings. However, the linguistic similarity between Dutch and German might have also affected the results.

Major (2007) compared the ratings of foreign accent in spoken Brazilian Portuguese by four rater groups, namely, native Brazilian Portuguese raters, native English raters who studied Brazilian Portuguese (Portuguese experience), native English raters who had not studied Portuguese (no Portuguese experience), and L2 English raters who had not studied Portuguese. The results showed that regardless of the rater's L1 background or familiarity with the L2, the four groups' ratings were highly correlated, suggesting that the L1 background and/or L2 familiarity did not affect the ratings of foreign accent.

Similarly, Sales (2012) examined the influence of raters' familiarity with accented speech on the degree of perceived foreign accent. In the Sales' study, native English raters were divided into two groups. The first group included teachers of English as a second language (ESL) and linguists. The second group included raters with no ESL experience nor linguistic training. Raters were asked to rate the degree of foreign accent in English speech produced by nonnative speakers of English. Their L1 backgrounds were Spanish, Arabic, Mandarin, Japanese, Korean, and Taiwanese. The results showed that raters with ESL experience or linguistics training rated foreign-accented speech more permissively (i.e., rated speech to be less accented) than those with no experience or linguistics training. The degree of familiarity with foreign accent based on their exposure to accented speech at the workplace or their living environment was correlated with their ratings. That is, raters with more ESL experience or advanced linguistics training were more lenient (i.e., rated speech to have less foreign accent) than raters with less ESL experience or linguistics training.

The proficiency level of nonnative raters has also been found to influence their rating of foreign-accented speech (e.g., Hayes-Harb et al., 2008; Xie & Fowler, 2013). Findings from the effect of proficiency generally have shown that nonnative raters with higher proficiency levels in the target language have a better perception of foreign-accented speech in the target language than intermediate L2 raters. Additionally, when the rater is familiar with the speaker's L1, this may facilitate better identification of transferred features from that L1 and better judgment of the accented speech.

On the role of proficiency and L1 use, Simonet (2010) examined the ratings of perceived foreign accent in the speech of highly proficient Spanish-Catalan early bilinguals by raters who were also highly proficient early bilinguals of Spanish-Catalan. The speakers and raters were divided into two subgroups based on their self-reported language dominance (Spanish-dominant vs. Catalan-dominant). Results revealed that Catalan-dominant raters were more successfully able to discriminate between Catalan-dominant and Spanish-dominant bilinguals speaking Catalan. Similarly, Spanish-dominant raters were more successfully able to discriminate between Spanish-dominant and Catalan-dominant bilinguals speaking Spanish.

Finally, Huang and Jun (2015) examined the correlation between rater's experience with accented speech, language proficiency, and the degree of perceived foreign accent in English speech produced by native and nonnative speakers (L1 Mandarin). The study had three rater groups: L1 English raters experienced in ESL teaching, inexperienced native English raters, and advanced L2 English raters who were enrolled in university classes and had eight years of mean age of arrival. The results showed that experienced native English raters were better at detecting native and nonnative speakers, as reflected by their foreign-accented speech ratings on a rating scale. Further, experienced native English and advanced nonnative raters were more lenient in their ratings of the degree of foreign accent (i.e., rating nonnative speech to be less accented) than the inexperienced native English group. However, the nonnative raters had lower inter-rater reliability scores because they were more variable in their ratings than both native raters' groups, whose scores were consistent.

To summarize, the divergent findings from previous studies demonstrate that the effects of rater-specific characteristics on the rating of foreign-accented speech are still not well understood. Although previous research has highlighted several factors related to the rater's background that contribute to their rating of global foreign accent, there are many questions raised by these studies that still need to be addressed. For example, there does not seem to be a consensus in the literature on which aspects of the rater's background are important to the ratings. For example, several studies have reported contradictory results regarding the rater's experience and familiarity with accented speech. Thus, it is unclear whether and how a rater's experience or familiarity with the accented speech provides any advantage when rating accented speech.

Another critical issue is that most of the previous research has not sufficiently considered is nonnative raters' inclusion. Only a few studies have included both native and nonnative raters (e.g., Flege, 1984; MacKay et al., 2006; Huang & Jun 2015). Therefore, there is undoubtedly more to be learned about whether nonnative raters could potentially provide comparable ratings of foreign-accented speech as native raters.

Additionally, there is an open question regarding the nonnative rater's L1 background and whether it influences the rating of accented speech in the target language (especially when the speaker is from the same L1 as the rater). This last issue has been addressed in a recent study by Almohareb (2016), which investigated whether naïve native and nonnative ratings of nonnative English speech are the same and whether the L1 background of the nonnative raters affects their ratings of English accented speech. The question is whether nonnative speakers of particular L2 rate speakers producing speech in this L2 differently depending on whether the rater shares the speaker's L1 or not.

This issue was explored in Almohareb (2016), where native and nonnative raters from different L1 backgrounds (English raters, Arabic raters, and Korean raters) rated the degree of global foreign accent on English phrases produced by native and nonnative speakers from different L1 backgrounds and different levels of proficiency (native English, advanced Korean, advanced Arabic, intermediate Korean and intermediate Arabic) using a 9-point rating scale. The results showed that both native and nonnative raters rank-ordered the speaker groups similarly. The results also revealed that native and nonnative raters rated speech samples produced by native English speakers to have the least degree of foreign accent. However, both nonnative rater groups (Arabic and Korean raters) diverged from the native English raters in terms of the degree of severity of the rating within each speaker group. The Arabic raters were generally more exacting in their ratings by assigning higher degrees of foreign accent to all of the nonnative speech samples, and the Korean raters were generally more lenient, by assigning lower degrees of foreign accent to the nonnative speech samples. Although these findings may suggest that perhaps the raters' L1 background does not directly influence their ratings, the number of raters (10 per group) might have been too small to draw reliable conclusions.

Another concern to raise with previous research is that no study has examined the degree of accentedness of the nonnative raters and whether their degrees of foreign accentedness correlate with their ratings of others' nonnative speech.

Finally, research needs to be clear about selecting nonnative speech samples and the speakers' proficiency levels. Although sample selection can be biased by the researcher’s idea of foreign accent, which might, in turn, affect the proficiency levels assigned to speakers, there are available measures that are less subjective that may be employed. For example, a panel of trained phoneticians may be used to judge a set of speech samples, creating a pool from which the research can choose speech samples based on clearly stated criteria.

To sum up, there remain many open questions to be answered regarding the relationship between the rater’s background and their global foreign accent ratings. In particular, exploring questions about the severity of differences between native and nonnative raters and between nonnative raters from different L1 backgrounds could have important implications for theoretical and methodological considerations. Theoretically, examining the differences between native and nonnative raters could contribute to our understanding of second language acquisition and how nonnative raters judge accented speech. Methodologically, including native and nonnative raters from different L1 backgrounds can help determine whether it is meaningful to include nonnative raters in future rating studies. Further, examining the relationship between the raters’ language background and their accented speech ratings could have important implications for high-stake speech assessment test research, including exploring potential rater biases and how to address them.

To this end, the current study aims to address previous issues by 1) including native raters and nonnative raters from different L1 backgrounds, 2) including English speech samples by speakers from different language backgrounds and varying degrees of foreign accentedness, and 3) collecting speech samples from the nonnative raters whose own speech is rated on a scale of foreign accentedness.

2.4.4. Research Questions and Hypotheses

The dissertation thus asks the following questions:

1. Do trained native English raters differ from un-trained native English raters in their rating of foreign-accented speech?

It is predicted that trained native raters and un-trained native raters will rate accented speech differently if experience (in the form of linguistic training) affects the trained native raters' judgments. However, it is also possible that both native raters' groups will rate accented speech similarly.

2. Do native and nonnative raters differ in their accent ratings of nonnative speech?
    a. How reliable are the accentedness ratings of nonnative raters compared to native raters?

For this question, there are three possibilities 1) native and nonnative raters will not differ in their rating of foreign-accented speech, 2) native and nonnative raters will differ from each other in rating foreign-accented speech, 3) differences between native raters and nonnative raters will be determined by the interaction with whether the nonnative rater shared an L1 with the nonnative speaker of not.

3. Regarding the degree of accentedness, do nonnative raters rate speech samples from the same L1 background as them differently from nonnative speech samples from different L1 backgrounds?

It is predicted that nonnative raters will rate speech from speakers who share the same L1 differently from nonnative samples from another L1 background. This is attributed to the raters' experience with their L1 accented speech in the target language and their knowledge of the speech in L1. However, the direction of the effect could be manifested in different ways. It is possible that nonnative raters rate speech from speakers with whom they share an L1 more leniently (compared to the other groups) by assigning lower accent rating compared to the other raters and at the same time rate speech from nonnative speaker with other L1 backgrounds in a more discriminating way by assigning higher accent ratings than other raters. It is also possible that they will only assign higher accent ratings to nonnative speakers from other L1 backgrounds without assigning lower accent ratings to the speakers with whom they share an L1 and vice versa.

4. Do other nonnative raters' characteristics, namely, raters' degree of accentedness, self-reported L1 use, and length of residence, influence their ratings of nonnative speech?

Although these factors' relationships with the ratings have not been widely tested before, it is possible to make some general predictions. If the rating of foreign accent is mediated or affected by these factors, then it would be predicted that these factors will influence the raters' ratings of accented speech. However, if the process of judging foreign accent is not mediated nor affected by the raters' degree of foreign accent, amount of L1 use, or length of residence, then it is predicted that these factors will not predict their ratings of accented speech.

Chapter 3

Methodology

This chapter describes the methodology, and data collection procedures for the rating experiment carried out in this dissertation. It describes the criteria for the stimuli selection for the foreign accent rating experiment, the online survey design and the requirement, and data collection procedures. In the rating experiment, 60 raters from three L1 backgrounds (20 Saudi Arabic L1/English L2, 20 Mandarin L1/English L2, and 20 native English) were asked to rate the degree of foreign accent on short English phrases produced by speakers from three different L1 backgrounds, namely English, Arabic and Mandarin.

3.1. Stimuli

The speech samples used in this foreign accent rating experiment were taken from the *Speech Accent Archive* (Weinberger, 2019). This archive is an online database of thousands of recordings by native and nonnative English speakers from various L1 backgrounds reading the same English passage, commonly known as *the Stella passage* (see Appendix A).

As an initial step in the nonnative samples' selection, recordings from 20 speakers from Saudi Arabic L1 background and 20 speakers from Mandarin Chinese L1 background were selected based on geographical region. Only speakers from the same regions were selected for each L1 background group (e.g., the central region of Saudi Arabia and the northern region of China). These 40 nonnative speech samples served as a pool for the final selection.

These 40 nonnative samples were uploaded to an online survey to collect judgments on the degree of foreign accent of these samples by trained phoneticians. The survey was sent to an advisory panel of five trained phoneticians. These trained raters were all native English speakers with graduate-level training in phonetics and ESL teaching from George Mason University. The trained raters independently listened to the entire passage and rated all 40 nonnative samples on the degree of foreign accentedness using a 9-point rating scale. In other words, the five trained raters listened to and rated the same 40 nonnative samples. The trained raters were not informed of the L1 backgrounds of the speakers. In the survey, all 40 samples were set to randomize with each run. They listened to the entire speech sample and used a sliding scale to give their rating. Accordingly, each sample received five ratings on the degree of foreign accent.

An interrater reliability test was performed by calculating average absolute agreement between trained raters using a two-way mixed intraclass correlation (ICC). The purpose of this test was to establish reliability scores for the ratings by the trained raters. The reliability score was computed using the Intraclass Correlation Coefficient (ICC) in the irr package (Gamer, Lemon, Fellows, & Singh, 2019) for R (R Development Core Team, 2014). The ICC score was .93 (95% CI [.87, .96]), indicating high interrater reliability. Accordingly, each speech sample was assigned a foreign accent score based on the advisory panel's average ratings. The trained phoneticians' ratings of the 40 nonnative speech created a pool of rated samples representing different degrees of foreign accent.

From the 40 nonnative samples, ten samples from L1 speakers of Mandarin Chinese and ten samples from L1 speakers of Saudi Arabic were selected to be used in the main rating experiment. The selection of these 20 samples was based on the following two considerations: 1) samples needed to be categorized into intermediate and advanced proficiency groups for each language background. Accordingly, samples with average ratings of 4 or less on the foreign accent scale were considered "advanced" and samples with average ratings of 5.5 or more were considered intermediate, as can be seen in Table 2. 2) within each proficiency group there needs to be some variation in the degree of accentedness. Therefore, each proficiency level had speakers with varying degrees of foreign accent. The advanced Arabic group had five speakers, and the intermediate Arabic group had five speakers. Similarly, the Mandarin advanced group had five speakers, and the Mandarin intermediate group had five speakers.

Further, the speakers from both languages in both proficiency level groups had similar degrees of foreign accentedness ratings. For example, when a Mandarin speaker who received an average foreign accent rating of 3 was selected to serve in the advanced Mandarin group, an Arabic speaker with the same average foreign accent rating of 3 was also selected for the advanced Arabic group. It is noteworthy that the matching between the speakers from the two L1 backgrounds based on their average foreign accentedness scores was only done for consistency in the design and did not factor in the analysis.

Accordingly, samples with similar degrees of accentedness between the two L1 backgrounds were selected.

Table 2 below shows the demographic information for the speakers from different L1 backgrounds. This demographic information was taken from *the Speech Accent Archive,* provided by the speakers when they originally submitted their speech samples to the database. In Table 2, age, age of onset, length of residence, and foreign accent rating were averaged across speakers in their respective groups with the ranges provided in parentheses. The term *age of onset* refers to the age of first exposure to English; the term *length of residence* refers to the duration of time the speaker spent in an English-speaking country (in this case the United States); the term *accent rating* refers to the average foreign accent rating provided by the trained raters.

Table 2
Speakers' Demographic information

Demographic variable	English (control)	Mandarin Chinese (advanced proficiency in English)	Mandarin Chinese (intermediate proficiency in English)	Saudi Arabic (advanced proficiency in English	Saudi Arabic (intermediate proficiency in English)
Mean Age	29.25 (18-53)	24.4 (18-29)	27 (21-31)	23.6 (19-30)	33.4 (18-57)
Gender	7m, 3f	2m, 3f	4m, 1f	2m, 3f	3m, 2f
Mean length of residence (yrs.)	28.75 (18-53)	4.7 (.5-10)	2.8 (.8-4.3)	1.3 (.1-3)	3 (1-8)
Mean Age of Onset (yrs.)	0	10.2 (6-15)	9.6 (6-12)	9.4 (6-15)	15.5 (12-21)
Mean accent rating (scale of 9)	1	3.3 (2.4 – 4)	6.6 (5.6 -8.4)	3.2 (2.4-4)	6.6 (5.8-8.4)

In addition to the 20 nonnative speakers, ten native American English speakers were selected from the archive to serve as a control group. The English native speakers were all from the state of Ohio. The reason for selecting speakers from the same state was for uniformity. The demographic information for native speakers is included in Table 2.

The same five phrases were selected from the Stella passage from each of the 30 speech samples (10 Arabic L1, 10 Mandarin L1, 10 English L1). The phrases were selected because they did not include any disfluencies (e.g., filled pause, false-start, self-correction) by any of the speakers. This yielded 150 target phrases (30 speakers x 5 phrases) used in the main online rating experiment. The phrases are listed in (1-5) below:

(1) Six spoons of fresh snow peas.

(2) Five thick slabs of blue cheese

(3) A big toy frog for the kids

(4) She can scoop these things into three red bags

(5) We will go meet her Wednesday at the train station

Each phrase of these five phrases was produced once by each of the 30 speakers resulting in 30 unique productions of each phrase.

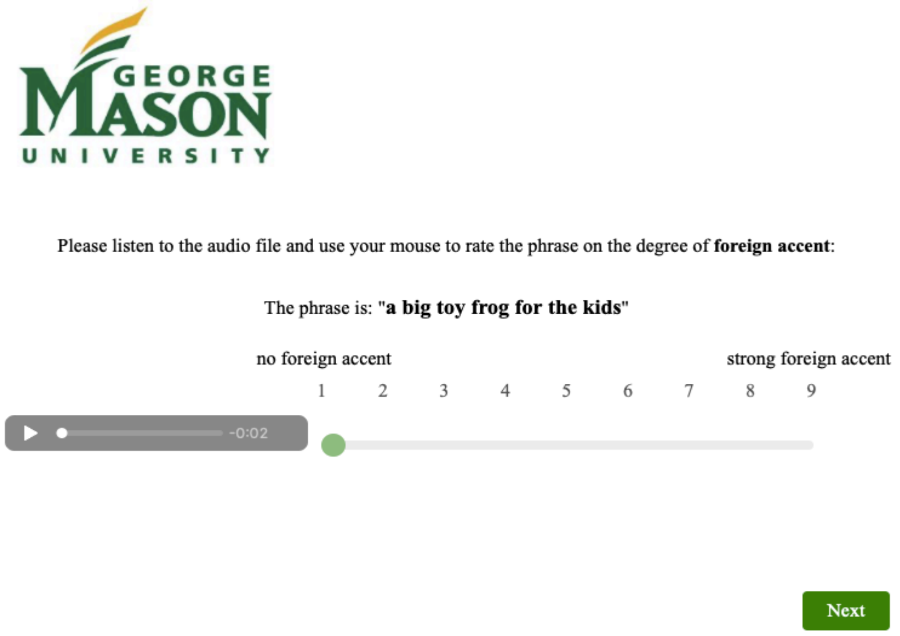
In addition to the 150 experimental phrases, six phrases were selected for the training session. The training phrases were not produced by the same speakers in the main experiment. The training phrases included one phrase produced by a native English speaker who produced the phrase "six spoons of fresh snow peas," and the other five training phrases were produced by nonnative speakers from a different L1 background (Turkish, Spanish, Russian, Italian, Korean); namely, each of the nonnative speakers produced one phrase of the five phrases listed above. The nonnative speech samples in the training session represented different degrees of foreign accents as determined by a trained phonetician.

3.2. Experiment Design and Procedure

The data collection took place in the Linguistics Program's acoustics lab at George Mason University. Participants (raters) were first asked to fill in a demographic information questionnaire (see Appendix B). After that, participants were asked to read the *Stella passage* and were audio-recorded, using the commercial version of iTalk application with Apogee 96K microphone plugged into an iPhone XS. Recordings were done at a sampling rate of 44.1k, 16-bit mono. The recordings of the raters were collected for further analysis, as will be described below. After the recording, participants were seated in front of a computer screen (MacBook Air) and put on headphones (Beats Studio wireless headphones) to take the online rating experiment.

The rating experiment was implemented in Qualtrics. The survey started with a welcome message that included general information about the study and the rating process's instructions. Participants were informed that they would listen to 156 English phrases produced by different speakers and that for each phrase, they needed to give a rating on a 9-point scale on the degree of the foreign accent. They were informed that they would take a short training session with six phrases and then start the main experiment, which was divided into two blocks with 75 samples in each block. The samples were randomly assigned to each block and randomized within the block with every run. This is done in order to minimize any possible order effects. The training session and the experiment had an identical design, and each phrase was presented on a separate page.

On the rating page for each sample (see Figure 1), participants saw radio buttons for the audio file and a written version of the phrase in that file along with the rating scale and the following statement: *Please listen to the audio file and use the mouse to rate it on the degree of foreign accent."* Scale point number 1 was located on the left end of the scale, and the phrase *"no foreign accent"* was written above it. Scale point number 9 was located on the right end of the scale, and the phrase *"strong foreign accent"* was written above it. The numbers from 2-8 marked the different scale points in between, with no statements written above them. Participants were asked to use the mouse to provide their rating and click the *Next* button to move on to the next sample. The audio files were set to play automatically. Participants had to provide a rating for each sample to move on to the next trial, and they were able to listen to the audio file more than once. However, participants were not allowed to go back to a previous sample once they moved on. The online rating experiment took approximately 25 minutes to complete, and participants were compensated with $10 upon the completion of the survey.



Please listen to the audio file and use your mouse to rate the phrase on the degree of **foreign accent**:

The phrase is: "**a big toy frog for the kids**"

no foreign accent								strong foreign accent
1	2	3	4	5	6	7	8	9

Figure 1 Screenshot of the Qualtrics' experiment interface

3.3. Participants

Sixty naïve raters with no reported speaking or hearing disorders were recruited to participate in the main rating experiment through advertisement flyers. The raters represented three different L1 backgrounds, namely, 20 native American English speakers, 20 native Mandarin Chinese speakers, and 20 native Saudi Arabic speakers. The raters were all students at George Mason University.

Table 3 below shows the demographic information for the raters. Table 3 represents their age, age of onset, length of residence, and L1 use and averages across raters.

Table 3
Raters' demographic information

Demographic variable	English (control)	Mandarin Chinese	Saudi Arabic
Mean Age	21.4 (18-33)	22.4 (19-27)	24.4 (18-35)
Gender	5m, 15f	6m, 14f	11m, 9f
Mean length of residence (yrs.)	N/A	2 (.1-10)	3.5 (.1-8)
Mean Age of Onset (yrs.)	0	8 (6-14)	12 (6-19)
L1 Use	100%	62% (40 -80)	58% (30 -90)
Mean accent rating (scale of 9)	1	6.14 (4- 8)	5.16 (1.6- 8.17)

In addition to the 60 raters who participated in the main rating experiment, five trained native English raters were recruited to provide accent ratings for the speech samples collected from the nonnative raters in the main experiment described above. The average accent ratings for the nonnative raters are provided in Table 3 above.

3.4. Evaluation of Nonnative Raters' Foreign Accents

The nonnative raters were recorded reading *the Stella passage* as part of their participation in this rating study. These recordings were collected to examine whether there is a relationship between the rater's degree of foreign accent and their accented speech ratings. For example, this process might help us explore whether a nonnative rater who rated nonnative speech samples is biased by their belief about their accent. If one of the raters had a strong foreign accent, would they treat any other accent as less accented because the speakers might seem to have a less pronounced foreign accent. In addition, could it be that nonnative raters can judge degrees of foreign accent at different levels independently from their foreign accent? Specifically, this analysis explores if there is a correlation between the nonnative ratings or nonnative samples and their degree of foreign accent as determined by trained phoneticians. This possible correlation is especially relevant if native and nonnative raters turn out to be different in their ratings.

The nonnative raters described in the participants' section were recorded, and their recordings were rated by the five trained native English phoneticians on the degree of foreign accentedness in an online survey. The survey employed the same methods described in the initial sample selection described above. The complete speech samples were uploaded in a Qualtrics survey, and each of the five trained raters listened independently to the entire speech sample and provided accent ratings. Accordingly, the five trained phoneticians' ratings were averaged for each speech sample, and the average served as the degree of accentedness for the nonnative raters.

Chapter 4

Results

4.1. Introduction

This chapter describes the statistical analysis of the data collected from the rating experiment described in Chapter 3 and reports on the findings. The purpose of each statistical analysis is described in its respective section. A total of 9000 ratings were collected and used in the analysis of the results. These ratings were obtained from 60 raters representing 3 L1 backgrounds. The raters were 20 native English raters, 20 L2 English raters from an Arabic L1 background, and 20 L2 English raters from a Mandarin L1 background. In this chapter, the rater groups are referred to as English raters, Arabic raters, and Mandarin raters. Each rater rated 150 English phrases produced by 30 speakers (5 phrases each) representing speakers from 3 L1 backgrounds with ten speakers. The speakers were native English speakers and nonnative English speakers from Arabic and Mandarin L1 backgrounds. The speaker groups are referred to as English speakers, Arabic speakers, and Mandarin speakers.

4.2. Interrater Reliability

Since this dissertation is concerned with the rater's overall foreign accent rating of the speech samples and how raters from the same group perform as a group, it was important to examine whether raters within their respective groups were consistent. This is typically measured by performing a reliability test, such as the Intraclass Correlation Coefficient (ICC). ICC is a widely used measure to establish the reliability of rating data collected from raters representing a group (see Koo & Li, 2016). Accordingly, interrater reliability tests were performed for each of the three rater groups by calculating average absolute agreement between raters using a two-way mixed intraclass correlation (ICC). The purpose of this test was to establish reliability scores for the rating data and groups. The reliability score was computed using the Intraclass Correlation Coefficient (ICC) in the irr package (Gamer, Lemon, Fellows, & Singh, 2019) for R (R Development Core Team, 2014).

The ICC scores were .98 (95% CI [.98, .99]) for the English raters, .97 (95% CI [.96, .98]) for Arabic raters, and .96 (95% CI [.96, .97]) for Mandarin raters. Typically, ICC values less than .5 indicate poor reliability, values between .5 and .75 indicate moderate reliability, values between .75 and .90 indicate good reliability and values greater than .90 generally indicate excellent reliability. In the current study, the ICCs were all greater than .95. This indicates that raters within their respective rater group rated each speech sample's degree of foreign accentedness similarly and consistently. The ICC values in this study are comparable to the values reported in foreign accent studies that used ICCs as an interrater reliability index (e.g., Huang & Jun 2015; Derwing & Munro, 2013; MacKay et al., 2006; Munro et al., 2006; Trofimovich, Lightbown, Halter, & Song, 2009).

4.3. Comparison Between Trained Native Raters and Un-trained Native Raters

This section explores the differences between trained and un-trained native English raters. To explore whether the foreign accent ratings were different between native raters with linguistic training and native raters without linguistic training, the descriptive statistics of how each of the native raters' group rated speech based on the speakers' L1 background are explored. The trained raters' rating data came from the trained phoneticians' responses collected in the initial stage of this experiment design, as described in Chapter 3. The data from the un-trained native raters came from the native English raters who participated in the experiment. In this section, the trained phonetician group is referred to as "trained raters," and the un-trained raters are referred to as "naïve raters."

The descriptive statistics showed that speech samples produced by native English speakers were rated as having almost the same degree of foreign accentedness by both native groups (naïve raters [M= 1.08, SD= 0.31] vs. trained raters [M= 1, SD=0].) However, trained raters were slightly better at detecting native speech than naïve raters. Speech samples produced by nonnative speakers with an Arabic L1 were rated as having similar degrees of foreign accentedness by naïve raters [M= 5.03, SD= 2.31] and trained raters [M= 4.90, SD=1.94]. Finally, speech samples produced by nonnative speakers with a Mandarin L1 were also rated as having similar degrees of foreign accentedness by naïve raters [M= 4.78, SD= 2.53] and trained raters [M= 4.94, SD=1.87].

Although these two groups' overall averages were generally closer in nonnative speech ratings, the slightly higher standard deviation for the naïve raters indicated more variability among their judgments. Additionally, the lower standard deviation from the mean in the ratings from trained raters indicated that their ratings tended to be less variable and closer together. This suggests that although naïve and trained raters' ratings may eventually converge, native raters may tend to agree more on their ratings. Table 4 below shows the average foreign accent rating given to each speaker by the two groups of native English raters (trained and naïve)

Table 4
Average foreign accent ratings by trained and naïve native English raters

native English speech			nonnative speech Arabic L1			nonnative speech Mandarin L1		
ID	Naïve	Trained	ID	Naïve	Trained	ID	Naïve	Trained
e104	1.08	1	a108	3.7	3.6	m109	3.24	3
e118	1.14	1	a109	3.8	2.6	m122	5.66	5.6
e126	1.05	1	a111	2.43	2.4	m26	5.96	6.8
e163	1.14	1	a116	6.61	6.2	m27	2.32	3
e237	1.05	1	a128	6.32	6.8	m54	4.38	4.6
e326	1.01	1	a155	4.6	4.4	m63	5.98	6
e419	1.23	1	a166	5.89	5.8	m73	2.55	3.4
e437	1.04	1	a83	6.48	6	m76	3.92	2.4
e494	1.02	1	a91	7.35	8.4	m80	7.69	8.4
e59	1.05	1	a96	3.1	2.8	m90	6.14	6.2

To further test this relationship, Pearson's correlation coefficient test was computed for these two groups' ratings. This test revealed a strong positive association between the ratings of trained and naïve native raters [r= 0.83, p<0.001]. This positive relationship does not indicate causality but is merely meant to show that these raters seem to strongly follow a similar rating pattern. This indicates that the two groups of native raters rated the same speech sample in a very similar way, and their ratings often exactly or approximately agree. This significant positive correlation is plotted in Figure 2 below.

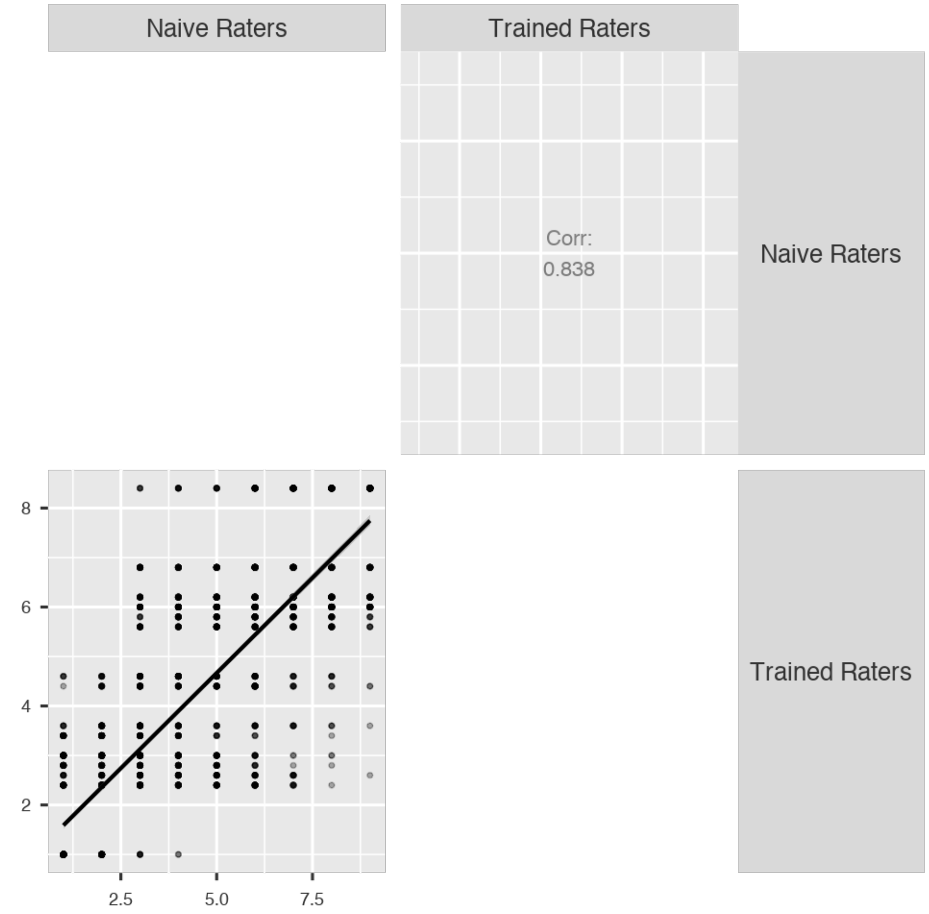


Figure 2 Correlation matrix for the foreign accent ratings between trained native English raters and naïve native English raters

To specifically test the differences between these two groups of native raters, a one-way repeated measures ANOVA was conducted to compare the effect of the L1 background of the speaker (English, Arabic, Mandarin) on the foreign accent rating. The results showed no significant effect of the rater group, F(1)=0.44, p= 0.503, indicating no overall difference between naïve and trained native raters.

However, a significant interaction between the rater group and the speaker L1 background was found: F(2)= 11.08, p <0.001. To further explore this interaction, post-hoc pairwise comparisons using Tukey's HSD tests implemented in *emmeans* package (Lenth, 2018) were carried out in the program R. The results of these tests revealed that although naïve and trained native English raters did not significantly differ in their ratings of speech produced by native English speakers (p >0.05) and nonnative speakers from Arabic L1 background (p >0.05), there was a statistically significant difference in their ratings of nonnative speech produced by Mandarin L1 speakers (p < 0.01) with a mean difference of 0.15 points on the rating scale. However, it is worth noting that this difference is minimal, so that we cannot conclude that these two rater groups were indeed different from each other. Figure 3 below illustrates how the two native English rater groups rated speech samples from speakers from the three different L1 backgrounds.

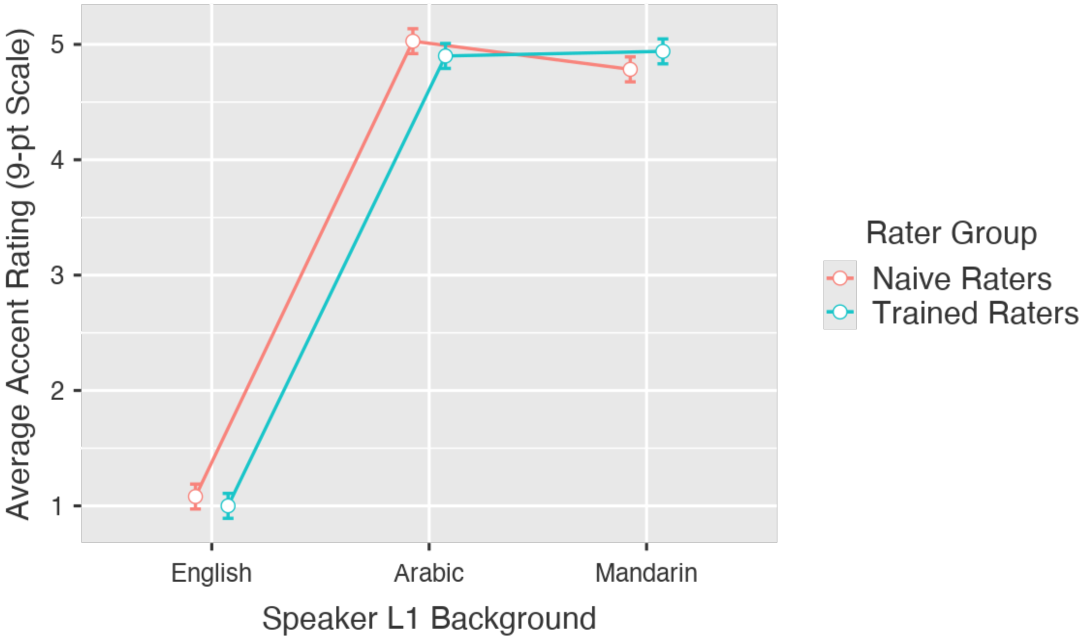


Figure 3 Average foreign accent ratings from naïve native English raters and trained native English raters (error bars represent 95% confidence interval).

4.4. Native versus Nonnative and Shared L1 versus Not-Shared L1

This section explores the effects of the raters' English status (native vs. nonnative) on their foreign accent rating. It also reports on the effects of the shared L1 status between the rater and the speaker (shared L1 vs. not-shared L1) on the foreign accent rating. This section starts by reporting the descriptive statistics and describing the observed patterns, then describes and discusses the findings from the statistical analysis.

First, in terms of the difference between native and nonnative raters, descriptive statistics revealed that, overall, native English raters assigned lower (=more English-like) accent ratings (M= 3.63, SD = 2.63) compared to the ratings from the nonnative raters (M= 4.22, SD= 2.77) for all the speech samples. Table 5 below shows the average accent rating and standard deviation for both rater's group (native vs. nonnative) divided by the speaker's L1. The magnitude of these differences is visualized in Figure 4.

Table 5
Average foreign accent ratings based on rater's English native-ness status

	Rater English Status			
	Native Raters		**Nonnative Raters**	
Speaker L1	Mean	*SD*	Mean	*SD*
English	1.08	0.314	1.73	1.24
Arabic	5.03	2.31	5.54	2.48
Mandarin	4.78	2.35	5.37	2.48

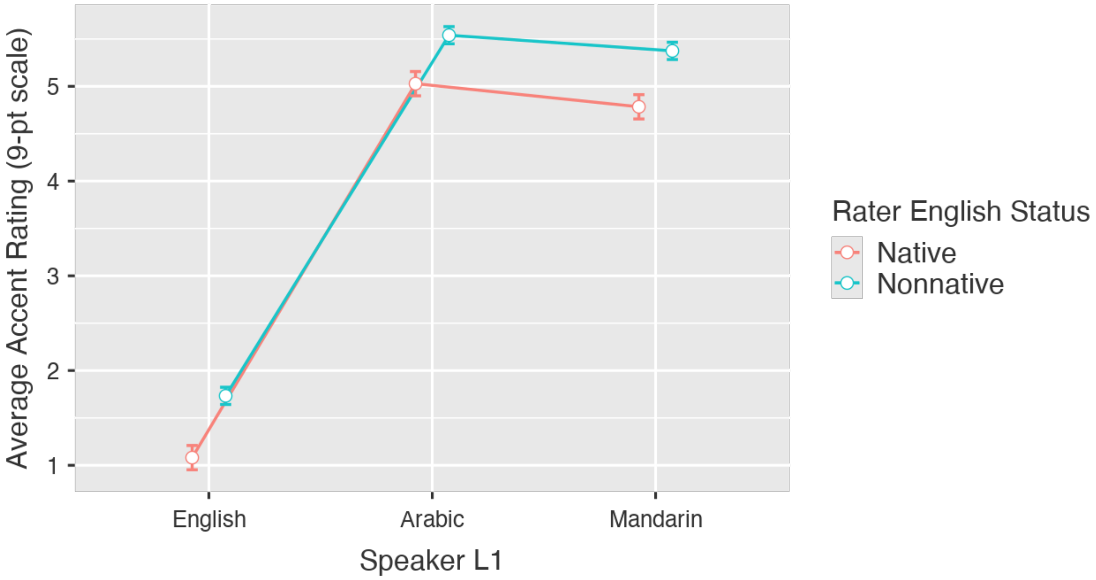


Figure 4 Average foreign accent ratings based on rater's English native-ness status (error bars represent 95% confidence interval).

As shown on Figure 4, there are differences in the average rating between native and nonnative raters when rating speech produced by English, Arabic, and Mandarin speakers. This pattern shows that nonnative raters typically assigned higher foreign accent ratings than native raters to all the speech sample groups.

In terms of the difference in the foreign accent rating between raters based on the shared L1 status between them and the speakers, initial descriptive statistics show that raters rated speech samples produced by speakers with whom they share an L1 to have less foreign accent (M=3.77, SD=2.76) than speech samples from speakers from a different L1 background (M=4.15, SD= 2.72). These differences are visualized in Figure 5 below, and the specific average ratings for each speaker and rater groups are summarized in Table 6.

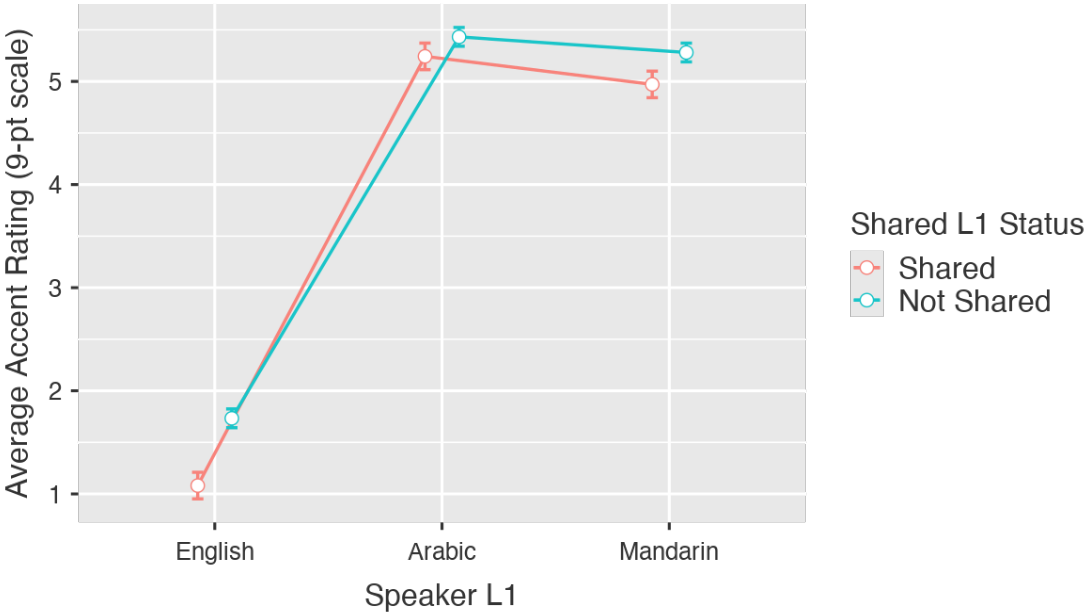


Figure 5 Average foreign accent ratings based on Rater- Speaker Shared L1 Status (shared L1 vs. not-shared L1) (error bars represent 95% confidence interval).

Table 6
Average foreign accent ratings based on Rater- Speaker Shared L1 Status (shared L1 vs. not-shared L1)

	Rater- Speaker Shared L1 Status			
	Shared L1		**Not Shared L1**	
Speaker L1	Mean	*SD*	Mean	*SD*
English	1.08	0.314	1.73	1.24
Arabic	5.24	2.41	5.43	2.45
Mandarin	4.97	2.47	5.28	2.43

To test these patterns, the accent ratings were modeled as the dependent variable in a mixed-effects regression model implemented in the program R (Baayen et al., 2008; R Core Team, 2014) using the lmer() function of the lme4 package (Bates, Maechler &

Bolker, 2013). Accordingly, models with different predictor variables and random effects were constructed and compared with likelihood ratio tests using the ANOVA function. Variables that significantly improved the model fit were retained in the best-fitting model. The variables for the rater's self-reported L1 use, rater's degree of accentedness, and rater's length of residence did not improve the model fit. The final model included two independent variables, namely, English rater status (native, nonnative), and speaker-rater shared L1 status (shared L1, not-shared L1), and the interaction between them. The model with maximal random effects structure failed to converge; therefore, a forward best-path method was used to determine which random slopes to include (Barr, Levy, Scheepers & Tily, 2013). Neither variable met the inclusion criterion (alph = .2), and thus the final model had minimal random effects structure with intercepts for participant and item. The parameter-specific *p*-values for the best fit model results were obtained by using the Satterthwaite approximation, implemented in the lmerTest package in R (Kuznetsova et al., 2017).

This statistical analysis aimed to determine if the English rater status (as being native or nonnative) and the speaker-rater shared L1 status (as being shared L1 or not shared L1) and the interaction between them have any effect on the foreign accentedness rating. It tested whether there was a statistically significant relationship between these raters' characteristics and their foreign accent rating. The best mixed-effects model results are shown in Table 7 below, with parameter estimate β statistics, standard error, *t* value, and *p*-value for the fixed effects.

Table 7

Coefficients of the best linear mixed-effects model of accent rating based on rater English status and speaker-rater shared L1 status (N= 9000).

		β	SE	*t*	*Pr(>|t|)*
Fixed effects					
(Intercept)		4.02	0.09	41.02	**<0.001**
Rater English Status (native)		-0.29	0.08	-3.48	**<0.001**
Shared L1 Status (same)		-0.19	0.02	-7.15	**<0.001**
Rater English Status* Shared L1 Status		-1.29	0.02	-45.42	**<0.001**
	s^2				
Random effects					
Participant (Intercept)	0.33				
Item (Intercept)	0.01				

As can be seen from Table 7, there was a significant main effect of the rater's English status on the foreign accent rating, indicating that native and nonnative raters were significantly different from each other. A significant main effect of speaker-rater shared L1 status was also found, indicating that raters were different when rating speakers from shared or not shared L1 background. Additionally, a significant interaction was found between the rater's English status and the speaker-rater shared L1 status, indicating differences among the three raters' groups.

To further explore the role of shared L1 background, the data was further divided by the rater L1 background to see how each rater's group performed. The raters' groups were divided based on their L1 background, mainly because when the ratings of speech samples produced by Arabic and Mandarin speakers were previously considered, the non-shared group had native and nonnative raters, which may not have provided a clear picture of what the pattern was. Accordingly, the average accent rating based on the rater

L1 background, and the speaker L1 background is presented in Table 8 and visualized in Figure 6 below.

Table 8
Average foreign accent ratings based on Rater and Speaker L1 background (English, Arabic, Mandarin)

	Rater L1 Background					
	English Raters		**Arabic Raters**		**Mandarin Raters**	
Speaker L1	Mean	*SD*	Mean	*SD*	Mean	*SD*
English	1.08	0.314	1.54	1.07	1.92	1.36
Arabic	5.03	2.31	5.24	2.41	5.84	2.51
Mandarin	4.78	2.35	5.78	2.41	4.97	2.47

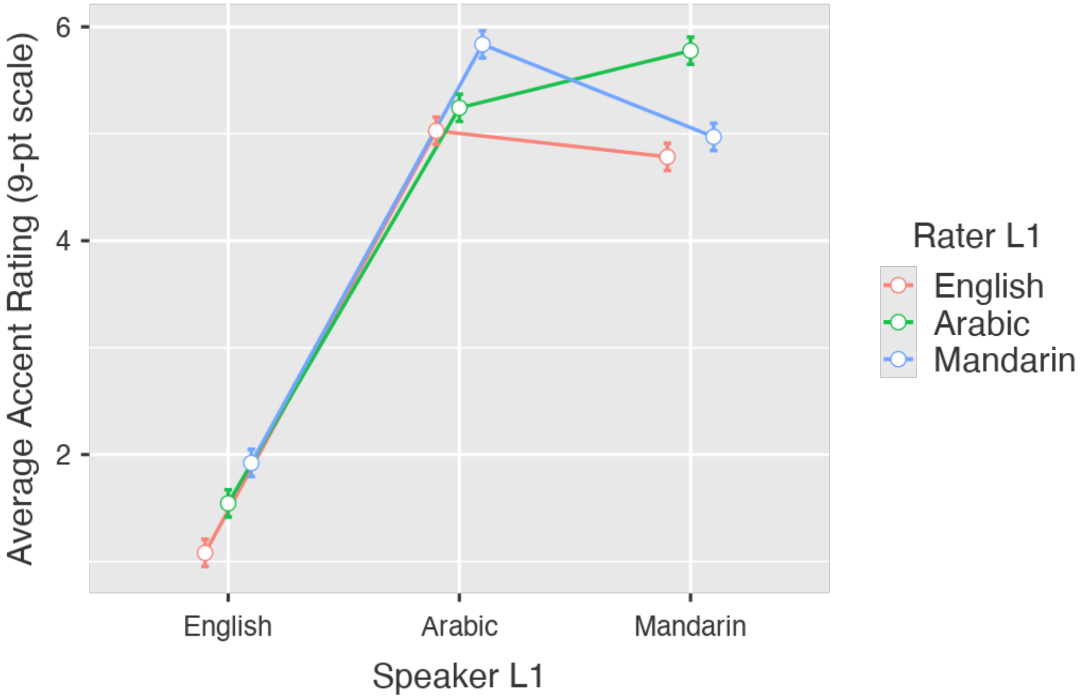


Figure 6 Average foreign accent ratings based on Rater and Speaker L1 background (English, Arabic, Mandarin) (error bars represent 95% confidence interval).

Figure 6 shows that native English raters always rated speech samples as having less foreign accent than the other two rater groups, regardless of the speaker group. Additionally, nonnative raters rated nonnative speech differently. Nonnative raters assigned lower foreign accent ratings to nonnative samples from speakers with whom they share the same L1 compared to the other raters and assigned higher foreign accent ratings to nonnative samples from speakers from a different L1 background compared to the other rater groups.

To further explore these differences' statistical significance, a model was constructed, but this time to test the interaction between the rater L1 background and the speaker L1 background. In this new model, the same method described above was followed. The purpose of this analysis was to specifically explore how raters from each L1 background rated speech from speakers with different L1 backgrounds. This new best fit model included the rating as a dependent variable and speaker L1 background and rater L1 background as the independent variables with intercepts for participant and item, as determined by a forward best-path method. The other variables (i.e., rater's L1 use, rater's length of residence, and rater's degree of accentedness) did not improve model fit and therefore were not retained in this final model. Model comparisons were then conducted to test the significance of these factors and their interaction by comparing the model with the best fit described here against the model without the effect in question. Accordingly, models with different predictors were computed and compared with likelihood ratio tests carried out by the ANOVA function, which uses log-likelihood to measure goodness of fit (cf. Baayen, 2008).

The results from the best mixed-effects model are shown in Table 9 below, with parameter estimate *β* statistics, standard error, t value, and *p*-value for the fixed effects.

Table 9
Coefficients of the best linear mixed-effects model of accent rating based on rater L1 and speaker L1 backgrounds (N= 9000).

		β	SE	*t*	*Pr(>|t|)*
Fixed effects					
(Intercept)		4.02	9.97	40.32	**<0.001**
Speaker L1 background		1.83	2.83	64.49	**<0.001**
Rater L1 background		2.78	8.97	3.10	**<0.01**
Speaker L1 * Rater L1		1.48	3.16	-4.70	**<0.001**
	s^2				
Random effects					
Participant (Intercept)	0.35				
Item (Intercept)	0.01				

As can be from the results in Table 9, there is a significant main effect of rater L1 background on the foreign accent rating, indicating differences between rater's groups. A significant main effect of speaker L1 background was also found, showing that speech samples from speakers from different L1 backgrounds were rated differently. However, more important is the interaction between these two factors; it reveals the nature of the differences in the ratings. In this regard, a significant interaction between rater L1 and speaker L1 backgrounds was found.

To further explore this interaction, post-hoc pairwise comparisons using Tukey's HSD tests implemented in *emmeans* package (Lenth, 2018) were conducted in the program R. The results of the post-hoc Tukey's HSD tests revealed that native English raters assigned significantly lower accent rating to speech produced by native English speaker than both Arabic and Mandarin raters (both p values <0.001). Arabic raters also assigned significantly lower ratings to the speech produced by English speakers than Mandarin raters (p <0.001). This finding indicates that the three raters' groups rated speech produced by native English speakers significantly differently. It also shows that English native raters are more accurate in their native speech ratings than the nonnative raters from Arabic and Mandarin L1 backgrounds, where Arabic raters were more accurate than Mandarin raters. That is, Arabic raters were closer in their rating of native English speakers to the native English raters.

The post-hoc tests further revealed that English and Arabic raters did not significantly differ in their speech samples' ratings produced by speakers from Arabic L1 backgrounds (p = .5). However, Mandarin raters assigned significantly higher foreign accent ratings to the speech samples produced by speakers from Arabic L1 backgrounds than both English and Arabic raters (both p values <0.001). Additionally, the post-hoc tests revealed that English and Mandarin raters did not significantly differ in their rating of speech samples produced by speakers from Mandarin L1 backgrounds (p =.3). However, Arabic raters assigned significantly higher foreign accent ratings to the speech samples produced by speakers from Mandarin L1 backgrounds than both English and Mandarin raters (p < 0.001). These findings indicate that nonnative raters were stricter in their foreign accent ratings of nonnative speech produced by speakers from a different L1 background (i.e., assigning higher foreign accent ratings). Overall, the findings from the statistical analyses suggest that, for the nonnative raters, sharing an L1 with the speaker does not necessarily imply assigning lower foreign accent ratings to those speakers, but rather nonnative raters assign higher foreign accent ratings to nonnative speakers from a different L1 background.

To explore which of the previous two models predicts the rating more accurately, a model comparison was conducted between the first model (Rater English Status* Shared L1 Status) and the second model (Speaker L1 * Rater L1). The model comparison revealed that the second model has a lower AIC score (AIC= 37903), indicating that it is the most parsimonious for the data. The interaction between the speaker and rater L1 backgrounds is a more accurate predictor of the ratings.

4.5. The Effects of Nonnative Rater's Degree of Foreign Accent, Self-Reported L1 Use, and Length of Residence on Their Ratings of Foreign Accent

This section explores whether other rater-specific factors affect their accent ratings. Specifically, it examines the nonnative raters' degree of accentedness, self-reported L1 use, and length of residence, and the shared L1 status in relation to their foreign accent ratings. Although the raters' degree of accentedness, self-reported L1 use, and length of residence in the United States did not improve the main mixed-effects regression models and were not included in the primary statistical analysis described above, in this section mixed-effects model tests are performed on a subset of the data, namely, the nonnative raters data. Three mixed-effects models were created to test these three effects following the same process described above for the primary analysis. In the best fit models, the foreign accent ratings served as the dependent variable, and the shared L1 status and factor in question (i.e., raters' degree of foreign accent, L1 use, and length of residence) were included as independent variables with item and subject as random effects and the shared L1 status as a slope.

The results from the best mixed-effects model for the raters' degree of foreign accent are shown in Table 10 below with parameter estimate *β* statistics, standard error, t value, and *p*-value for the fixed effects.

Table 10
Coefficients of the best linear mixed-effects model of accent rating based on the nonnative raters' accent and shared L1 status (N= 4000).

	β	SE	*t*	*Pr(>|t|)*
Fixed effects				
(Intercept)	5.28	0.22	23.51	**<0.001**
Raters' accent degree	-0.18	0.10	-1.77	0.08
Shared L1 status	-0.58	0.13	-4.39	**<0.001**
Raters' accent degree * Shared L1 Status	-0.08	0.07	-1.14	0.25
	s^2			
Random effects				
Participant (Intercept)	1.35			
Shared L1 status (slope)	0.48			
Item (Intercept)	0.07			

As shown in Table 10, there was a significant effect of the shared L1 status, indicating that nonnative raters rated nonnative speech produced by speakers with whom they share an L1 differently from when speakers were from a non-shared L1 background, as reported in section 4.4 above. The raters assign lower foreign accent ratings to speech produced by speakers with shared L1background with them and assign higher foreign accent ratings to speakers from a non-shared L1 background. However, neither the effect of raters' degree of foreign accent nor the interaction between of raters' degree of foreign accent and shared L1 status were significant, indicating the raters' own degree of foreign accent and its interaction with the shared L1 factor do not predict the foreign accent ratings in this data.

Similar findings were also found for the self-reported L1 use. The results from the best mixed-effects model for the raters' self-reported L1 use are shown in Table 11 below, with parameter estimate *β* statistics, standard error, t value, and *p*-value for the fixed effects.

Table 11
Coefficients of the best linear mixed-effects model of accent ratings based on the nonnative raters' self-reported L1 use and shared L1 status (N= 4000).

		β	SE	*t*	*Pr(>|t|)*
Fixed effects					
(Intercept)		5.28	0.22	23.55	**<0.001**
Raters' L1 use		-0.02	0.03	-1.83	0.07
Shared L1 status		-0.58	0.13	-4.38	**<0.001**
Raters' L1 use* Shared L1 Status		-0.01	0.01	-1.03	0.3
	s^2				
Random effects					
Participant (Intercept)	1.34				
Shared L1 status (slope)	0.48				
Item (Intercept)	0.07				

This model's findings indicate that the nonnative raters' L1 use and its interaction with the shared L1 status do not predict the foreign accent ratings provided by the nonnative raters.

The third model examined the raters' length of residence and the shared L1 status and the interaction between them in relation to the foreign accent ratings provided by nonnative raters. The results from the best mixed-effects model for the raters' degree of foreign accent are shown in Table 12 below with parameter estimate *β* statistics, standard error, t value, and *p*-value for the fixed effects.

Table 12
Coefficients of the best linear mixed-effects model of accent ratings based on the nonnative raters' length of residence and shared L1 status (N= 4000).

	β	SE	*t*	*Pr(>|t|)*
Fixed effects				
(Intercept)	5.28	0.22	23.81	**<0.001**
Raters' LoR	0.14	0.06	2.15	**<0.05**
Shared L1 status	-0.58	0.13	-4.35	**<0.001**
Raters' LoR* Shared L1 Status	-0.03	0.04	0.81	0.4
	s^2			
Random effects				
Participant (Intercept)	1.29			
Shared L1 status (slope)	0.49			
Item (Intercept)	0.07			

As can be seen from Table 12, the model revealed a significant main effect of raters' LoR on their ratings of foreign accent, indicating that raters with different lengths of residence rated accented speech differently. The analysis also revealed a significant interaction between raters' LoR and the shared L1 status, indicating that raters with different lengths of residence differed in their rating of accented speech based on whether the speech was produced by speakers sharing the same L1 background with them. This interaction is shown in Figure 7 below.

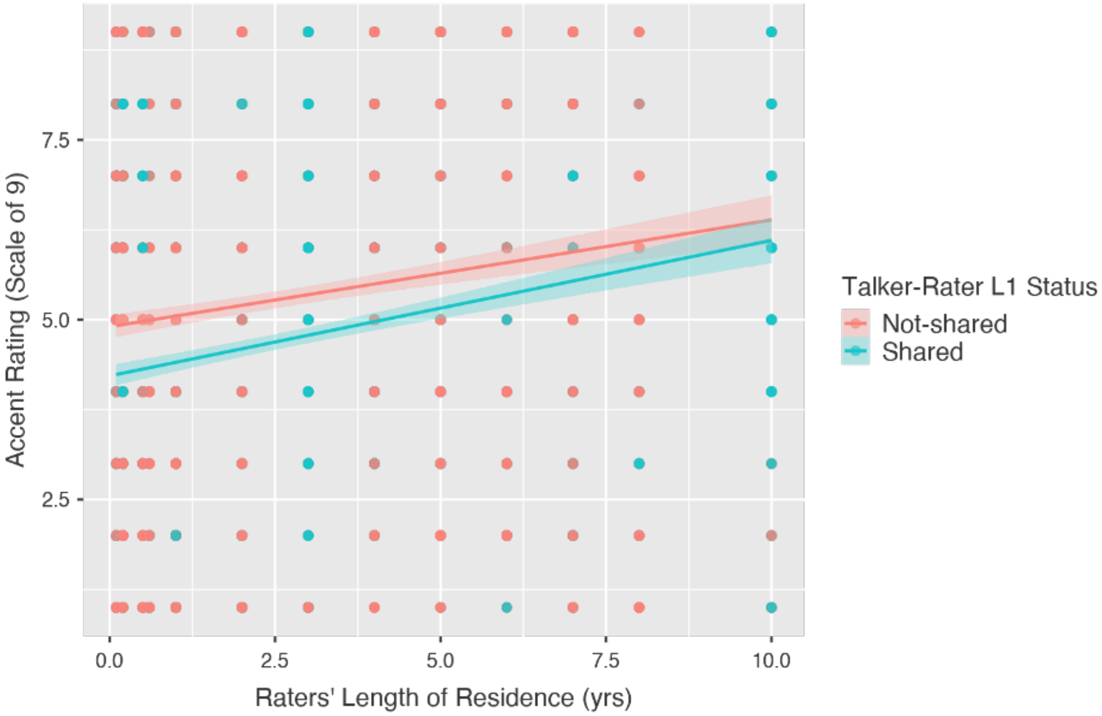


Figure 7 Scatterplot of the nonnative raters' accent ratings and their LoR based on the shared L1 status between the talker and rater (Ribbons represent 95% confidence interval around the accent ratings).

As shown in Figure 7, nonnative raters generally rated speech from speakers sharing the same L1 background as less accented and speech from speakers with a different L1 background to be more accented. However, regardless of the shared L1 status between the speaker and rater, nonnative raters with shorter length of residence typically assign lower foreign accent ratings than nonnative raters with longer lengths of residence in the United States. However, it is worth noting that the ratings assigned by nonnative raters with long lengths of residence seem to converge regardless of whether the speaker's L1 is shared with them.

Additionally, although it does not directly relate to the relationship to the nonnative accent ratings, multiple Pearson's product-moment correlation coefficient were computed to assess the relationship between the raters' degree of foreign accent, self-reported L1 use, and length of residence in the United States. The correlation tests revealed significant associations between the raters' degrees of accentedness and their self-reported L1, and their length of residence, and between raters' length of residence and self-reported L1 use. First, there was a weak positive correlation between the raters' degree of foreign accent and their self-reported L1 use [$r= 0.45$, $p < 0.01$]. This indicates that the raters who reported lower L1 use were rated to have relatively less foreign accent than raters who reported high L1 use. Second, there was a moderate negative correlation between raters' degree of foreign accent and their length of residence [$r= -0.61$, $p < 0.001$]. That is, raters who reported longer lengths of residence in the United States were rated to have less foreign accent, and raters who reported shorter lengths of residence were rated to have more foreign accent. Third, there was a weak negative correlation between raters' length of residence and their self-reported L1 use [$r= -0.34$, $p < 0.05$]. This indicates that rater's with longer length of residence tend to use their L1 less than raters with shorter lengths of residence. These correlations, though not strong, are informative for the development of L2 competence.

4.6. Summary of the Results

This chapter analyzed the data examining the effects of rater's specific characteristics on the foreign accent ratings obtained from 60 raters rating the same English speech samples. The raters were divided into three equal groups of 20 raters, each based on their L1 background. The L1 backgrounds were native English, nonnative English raters from Arabic L1 backgrounds, and nonnative English raters from Mandarin L1 backgrounds. The speech samples represent native and nonnative English speech produced by speakers from English, Arabic, and Mandarin L1 backgrounds. The total number of ratings considered in this chapter were 9000 ratings (60 raters x 150 speech samples).

As a first step in analyzing the data, Intraclass Correlation Coefficient tests were carried out to establish the three rater groups' reliability scores. The ICC tests revealed that all raters within their respective groups were highly consistent in their ratings, indicating that their ratings were reliable. The differences between trained native English raters and naïve native English raters were then examined to explore whether the trained raters' experience with accented speech resulted in different rating behavior than naïve raters. The analysis revealed that both groups' ratings were strongly correlated, which indicated a similar pattern of their ratings. Although the trained raters were slightly less variable in their ratings than naïve raters, both groups' ratings seem to converge at the end. The only statistically significant difference found between the two groups revealed that naïve raters assigned higher accent ratings to speech samples from speakers from Mandarin L1 background. However, this significant difference was minimal on the scale and perhaps not enough to set these two groups apart.

The data was then analyzed in terms of the rater's specific characteristics. For this analysis, descriptive statistics were first explored, which revealed differences between the raters' groups regarding the speaker's English status (native vs. nonnative) and the speaker-rater shared L1 status (shared L1 vs. not-shared L1). The patterns observed from the descriptive statistics indicate that native raters were more lenient than nonnative raters by assigning lower foreign accent ratings to all of the speech sample groups.

Additionally, it was found that raters were more lenient when rating speech from speakers with whom they share an L1 than when rating speech from speakers from a different L1 background than the rater. This is demonstrated by the lower accent ratings they assigned to the speech from the shared L1 background. However, the magnitude of the effect of the native and nonnative raters' distinction was greater than that of the shared vs. non-shared L1 distinction. Accordingly, the rater's L1 background was used as a grouping factor to see how raters from each rater group rate speech from the three speakers L1 backgrounds. This was done to explore the specific nature of the effect. This analysis revealed that nonnative raters did not necessarily assign lower accent ratings to speech samples from the shared L1, but instead, they seem to assign higher foreign accent ratings to nonnative speech from speakers who did not share an L1 background with them.

To further explore these patterns, multiple mixed-effects regression models with the different factors above were conducted. The significant finding from the mixed-effects regression models revealed that the rater's English status (native vs. nonnative) and the speaker-rater shared L1 status (shared L1 vs. not-shared L1) were significant predictors of their ratings, indicating that native and nonnative raters were different from each other in rating accented speech and that raters were different in their ratings of speech samples from speakers with whom they share an L1 and those with whom they do not share an L1. The results from the mixed-effects regression model testing the effect of the rater L1 background confirmed these effects and, more importantly, revealed that when nonnative raters rated nonnative speech from speakers with a shared L1, they did not significantly differ from native raters rating the same speech. However, when raters rated nonnative speech from a not-shared L1 background, they assigned higher foreign accent ratings to those samples. This demonstrates that nonnative raters do not assign lower accent ratings to nonnative speech samples from speakers with whom they share an L1, but instead, they assign higher accent ratings to nonnative speech samples from speakers from a different L1 background (i.e., not-shared L1).

Additional mixed-effects models were run on a subset of the data to explore the effects of the raters' degree of foreign, L1 use, and length of residence on their accented speech rating. The analyses revealed that raters' degree of foreign accent and L1 use do not predict the foreign accent ratings, nor do the interactions between these two factors and the shared L1 status between the speaker and rater. However, the length of residence factor and its interaction with the shared L1 status were found to affect the nonnative raters' ratings of accented speech. Specifically, nonnative raters with a shorter length of residence assigned lower foreign accent scores than nonnative raters with longer lengths of residence. Further, nonnative raters rated speech produced by speakers from the same L1 background as them to have less foreign accent than speech from speakers with a different L1 background. These findings and their implications for the field of foreign accent are discussed in Chapter 5.

Chapter 5

Discussion and Conclusion

5.1. Introduction

This dissertation aimed to examine the effect of rater-specific characteristics on the rating of foreign-accented speech. Four main questions were asked. These were related to the native rater's experience (trained vs. not trained), the rater's English status (native vs. nonnative), the speaker-rater shared L1 status (shared L1 vs. not-shared L1), and the relationship between nonnative raters' self-reported L1 use, length of residence in an English-speaking country, and own degree of foreign accent, and their ratings of accented speech. In this chapter, the implications of the findings are discussed as they relate to foreign accent research. Accordingly, each research question is addressed separately.

5.2. RQ1: Do Trained Native English Raters Differ from Un-Trained Native English Raters in Their Rating of Foreign-Accented Speech?

For this question, native English raters' ratings with linguistic training (particularly phonetics) and un-trained native English raters were compared to examine whether these two native rater groups differ from each other based on linguistic training. The analyses revealed that the ratings of both native rater groups were strongly correlated. Further, the pairwise comparisons revealed that they were generally similar in their ratings. The exception was that un-trained raters assigned higher accent ratings (more accented) to speech samples produced by speakers from Mandarin L1 background. However, given that the mean difference was minimal, it is argued that it is not enough to differentiate the two native rater groups reliably.

In previous research, the role of experience on the rating of accented speech (e.g., language teaching and linguistic training) have yielded divergent findings. For example, research by Thompson (1991) and Huang et al. (2015) have claimed that raters with no experience with accented speech were generally more exacting (i.e., assigned higher accent ratings) than experienced raters. Similar findings have also been reported in oral language assessment tests (e.g., Fayer & Krasinski, 1987; Hsieh, 2011). However, by contrast, Hadden (1991) found that inexperienced raters were more lenient (i.e., assigned lower accent ratings) than experienced raters.

Although this dissertation’s results do not necessarily support either of these two claims, it is suspected that perhaps prior exposure to accented speech rather than linguistic training might have contributed to the lack of apparent differences between trained and un-trained native raters. In particular, in this dissertation, raters from both native groups live in Northern Virginia and attend George Mason University. Northern Virginia is a multicultural and multilingual community, and George Mason University is voted Virginia’s most diverse university (US News & World Report, 2019), where many languages, including Arabic and Mandarin accented English, are widely spoken. Raters from both native groups were perhaps familiar with a fuller range of accented speech, which may explain the similarities found between them. Future research will need to control for other possible sources of experience.

5.3. RQ2: Do Native and Nonnative Raters Differ in Their Accent Ratings?

This research question examined native and nonnative raters' ratings of foreign-accented speech to determine whether these two rater groups' judgments would differ. First, it was found that native and nonnative rater groups were highly consistent in their rating of accented speech with high-reliability scores indicating that raters within their group were reliable. These reliability scores obtained from the ICC tests were comparable to those reported in previous accent studies employing the same measure (e.g., Huang & Jun 2015; Derwing & Munro, 2013; MacKay et al., 2006; Munro et al., 2006; Trofimovich, Lightbown, Halter, & Song, 2009). Establishing reliability scores was very important for the data because it has implications for the subsequent findings' validity and generalizability.

The current study found that native and nonnative raters differed in their native and nonnative speech accent rating. This finding is consistent with previous studies that found differences between native and nonnative raters (e.g., Flege, 1988; Major, 2007). The current analyses revealed that native raters rated speech from native and nonnative speakers significantly lower (less accented) than nonnative raters. In terms of speech samples from native English speakers, nonnative raters assigned higher accent scores (more accented) than native raters. This demonstrates that nonnative raters differed from native raters in judging native speech. In terms of nonnative speech, again, nonnative raters were stricter than native raters by assigning higher accent ratings, indicating that they detected more foreign accent than native raters did. These patterns are consistent with previous literature (e.g., Sheorey, 1985; Fayer & Krazinski, 1987; Caban, 2003; Kang, 2012; Schoonmaker-Gates, 2012).

There are several possible explanations for the difference between native and nonnative raters. A possible explanation could be related to linguistic competence in the target language. Nonnative raters do not have the same linguistic competence in the target language that native raters have. This is partly because nonnative raters learned the target language after they have fully acquired an L1, and thus other factors are affecting their competence. Additionally, the nonnative raters have different proficiency levels, which might have affected their ratings. Similar findings were reported by Scovel (1981), where he found that nonnative raters did not approximate the ratings of different groups of native raters even when the nonnative raters were advanced in the L2.

The current study also explored whether nonnative raters with longer lengths of residence, lower L1 use, and less foreign accent (as judged by trained raters) approximate native English raters' scoring behavior and whether those nonnative raters were different to other nonnative raters with shorter length or residence, higher L1 use, and more degrees of foreign accent. However, it was found that nonnative raters with longer lengths of residence, lower L1 use, and less degrees of foreign accent did not approach native raters, nor did their rating behavior differ significantly from the other nonnative raters with shorter length or residence, higher L1 use, and higher degrees foreign accent. This is perhaps because a combination of these factors (i.e., length of residence, L1 use, and rater's degree of accent) is not an accurate indication of proficiency. Instead, a more objective measure of proficiency is necessary to evaluate these possibilities. Accordingly, developing a proficiency score for the nonnative raters in this study may prove unfeasible without such a measure.

A more plausible explanation for the difference between the ratings obtained from native versus nonnative raters could be due to the strategies native and nonnative raters employ when rating accented speech. Previous research focusing on raters' feedback suggests that native and nonnative raters utilize different strategies when judging L2 speech. For example, Jun and Li (2010) reported that nonnative raters relied more on segmental errors and suprasegmental errors when rating the L2 speech whereas native raters focused on more general considerations such as clarity of speech and ease of understanding (also see Crowther, Trofimovich, & Isaacs; 2016). These different strategies could be attributed to the way native and nonnative speakers acquire the language. As Schoonmaker-Gates (2012) argues, while native speakers do not typically receive explicit pronunciation instruction of their native language, nonnative learners typically do as part of their academic language programs. Accordingly, if nonnative raters rely on more fine-grained acoustic cues when rating (presumably motivated by the pronunciation instruction), this perhaps explains why they might rate speech as more accented and consequently assign higher accent ratings than native raters do.

Another possible explanation for the difference observed between native and nonnative raters could be related to the nonnative rater's L1 background. This possibility is explored further in section 5.4 below.

5.4. RQ3: In Terms of The Degree of Accentedness, Do Nonnative Raters Rate Speech Samples from their L1 Background Differently from Nonnative Speech Samples from Different L1 Backgrounds?

This question examined whether nonnative raters from Arabic and Mandarin L1 backgrounds rated accented speech as less or more accented due to the shared L1 status. It was found that nonnative raters did not differ significantly from native raters when rating speech from nonnative speakers sharing the same L1 background with them (i.e., with the nonnative raters). In particular, nonnative raters from Arabic L1 backgrounds did not differ significantly from native English raters when rating nonnative speech from speakers of Arabic L1 background. Similarly, nonnative raters from Mandarin L1 backgrounds did not differ significantly from native English raters when rating nonnative speech from speakers of Mandarin L1 background.

These findings suggest that nonnative raters rate speech from speakers with whom they share an L1 as less accented than the other nonnative judges. However, nonnative raters rated nonnative speech produced by speakers from different L1 backgrounds to be more accented. In particular, Arabic raters rated accented speech from Mandarin L1 backgrounds to be more accented than English and Mandarin raters' ratings. Similarly, Mandarin raters rated accented speech from Arabic L1 background to be more accented than English and Arabic raters' ratings. Previous findings have sometimes reported that nonnative raters rated accented speech as less accented when the speaker and the rater share an L1 (e.g., Yuan et al., 2010), but other studies found the opposite (e.g., Schoonmaker-Gates; 2012). Besides, some studies have found that nonnative raters did not differ in their rating of accented speech according to their L1 background and whether it was shared or not shared with the speaker (e.g., Munro et al., 2006).

There are different possible explanations for the results obtained in the current study. First, nonnative raters may have less familiarity with the nonnative speech from a different L1 background and have difficulty processing it and thus rate it as more accented. This idea is supported by evidence from speech processing studies (Bradlow & Bent, 2003, 2008), which suggests that repeated exposure to a specific foreign accent leads to faster adaptation to the specific accent and facilitates better processing, comprehension, and intelligibility. Accordingly, because nonnative raters are familiar with accented speech by a speaker sharing the same L1 with them, they have perhaps adapted to that type of accent, making it easier for them to perceive and thus provide a more accurate accent rating. Conversely, because nonnative raters are not familiar or (are less familiar) with accented speech by speakers not sharing the same L1 with them, they may not have fully adapted to that type of accent, making it harder for them to process and thus rate it as more accented.

These findings from the current study seem to point to a possible bias by nonnative raters toward nonnative speech by speakers from a different L1 background. If this is the case, it has consequences and important implications for second language perception studies and spoken language assessment research and practice—these implications are discussed in section 5.6.

5.5. RQ4: Do Other Nonnative Factors Including Raters' Degree of Accentedness Self-Reported L1 Use and Length of Residence Influence Their Rating of Accented Speech?

This question explored whether other nonnative rater-specific characteristics influence their rating of accented speech. The first test explored whether a nonnative rater's degree of foreign accent (as determined by trained raters) predicts their accented speech ratings. The idea was to see if raters with higher degrees of foreign accent, for example, were more lenient or more scrupulous in their accent ratings and vice versa. To my knowledge, the possible relationship between a rater's degree of foreign accent and the rating they assign to accented speech has not been previously examined.

The results from this study indicate that the nonnative raters' degree of foreign accent does not significantly influence their rating of accented speech. In other words, nonnative raters with different degrees of foreign accent rated speech the same way. This is further supported by the high ICC scores for the groups. The fact that nonnative raters (even those with strong foreign accents) seem to perform better on the rating task than on the production task suggests that production and perception are not closely linked in this phenomenon. Thus, nonnative raters with a strong foreign accent may be able to judge foreign accentedness as equally as nonnative raters with less foreign accent.

There are different possible explanations for this result. One could be that nonnative raters are not aware of their accent but are aware of others' accents. Accordingly, their judgments of accents do not originate from their beliefs about their own foreign accent as the basis for their judgment. It is also possible that nonnative raters' perception ability exceeds their production ability (e.g., Felge, 1988). Nonnative raters know how native speakers sound and their judgments of foreign accents are determined by their perceived divergence from this internalized knowledge. However, nonnative raters' ability to produce these native-like forms may be affected by the restrictions associated with acquiring a second language. Therefore, a heavily accented nonnative rater's ability to judge foreign accent might be the same as another nonnative rater's ability with less foreign accent.

The second test evaluated the effects of nonnative raters' self-reported L1 use and their accent ratings. Although the amount of L1 use is a subjective measure, it provides an indirect insight into the rater's L2 experience. The amount of using one language is inversely related to the amount of using the other (i.e., the higher L1 use, the lower L2 use, and vice versa). Although previous studies have typically found that raters with more experience with the L2 show more native-like ratings of foreign accent, the amount of experience is generally determined by the rater's age, length of L2 residence, and familiarity with the accent. Much less is understood about the effect of L1 use. The current analysis results revealed that the raters' self-reported L1 use does not seem to influence their ratings of others' accents. This could be explained by the fact that most raters reported high percentages of daily L1 use.

The third test examined the effect of the raters' length of residence and their ratings of accented speech. Findings from previous research (e.g., Flege, 1988) have shown that nonnative raters with longer lengths of residence in a country perform differently from nonnative raters with shorter lengths of residence. The results revealed that raters with a shorter length of residence typically assign lower accent ratings than raters with longer residence lengths. It was also revealed that nonnative raters' ratings of speech samples differed based on whether the speech sample's speaker has a shared L1 with the rater. However, the interaction between the L1 status and length of residence revealed that regardless of the shared L1 status, nonnative raters with shorter LoR still assigned lower foreign accent scores than raters with longer LoR.

5.6. General Discussion and Implications

The current study results have revealed that native and nonnative raters judge the degree of foreign accent differently. The results also revealed interesting differences between nonnative raters from different L1 backgrounds when rating nonnative speech produced by nonnative speakers from various L1 backgrounds. These findings have important implications for research on the rating of foreign accent, our understanding of the development of L2 competence, language pedagogy, and the practice of spoken language assessment. Additionally, although the current data did not show that nonnative raters differ in their rating of accented speech due to other rater-specific characteristics such as L1 use and L2 length of residence, these factors may have further implications for these fields of research. In the following, the implications for four fields of research and practice are discussed.

5.6.1. The Rating of Foreign Accent

While most previous research on the rating of foreign accent typically used native raters as their optimal judges of accented speech, native raters are not representative of all types of raters. Accordingly, including nonnative raters in foreign accent rating studies will contribute to our understanding of the nature of foreign accent rating and provide valuable information about the development of L2 perception. The results from the current study and previous research that used nonnative raters indicate that differences exist between native and nonnative raters, motivating further research investigating these differences and their implications for foreign accent rating and related fields. This is particularly important since the nature of these differences seems to be interrelated with other factors inherent to the rater.

For example, in the current study, it was found that the incongruence in the L1 background between nonnative speakers and raters significantly contributed to this observed difference. Previous studies have also found other factors to be significant, such as the proficiency level of the nonnative rater, and amount of exposure to speech in the target language to be contributing factors to the difference between native and nonnative raters (e.g., Flege, 1988; Neufeld, 1980; Bradlow & Bent, 2003, 2008; Schoonmaker-Gates, 2012), although these factors were not found to be significant in the current study. Furthermore, evidence from foreign accent studies employing a think-aloud or feedback methodology (e.g., Jun & Li, 2010; Zhang & Elder, 2011; Crowther et al., 2016) suggests that native and nonnative raters may rely on different strategies when rating accented speech. These findings demonstrate that additional research is necessary to determine the nature of the differences between native and nonnative raters, and its interacting factors, and assess its further implications.

5.6.2. L2 Competence

While the current study did not find differences between nonnative raters within their groups based on their L1 use, length of residence, and degree of foreign accent, these factors could still be investigated in various ways in future research, especially with consideration to the overall factor of competence in the L2. This possibility is particularly supported by the differences reported in previous research, which examined some of these factors and found them to influence the accent ratings (e.g., Flege, 1988). However, the fact that there were differences between native and nonnative raters may well be due to the overall language competence. While native raters perhaps exclusively rely on their native competence to judge accented speech, it is not clear what nonnative raters rely on when judging accented speech.

Previous research has typically examined L2 competence in terms of the proficiency levels of the raters. These studies found that nonnative raters varied in their accented speech rating due to their L2 proficiency (e.g., Schoonmaker-Gates; 2012). Further, familiarity with a particular accent has been reported to facilitate a better understanding of that accent and thus influence the way it is rated. In the current data, nonnative raters rating speech from speakers sharing an L1 with them did not significantly differ from native raters rating the same speech. Further, previous research shows that unfamiliarity with a specific nonnative accent generally results in perceiving it as having a higher degree of foreign accent than a familiar accent, which extends to affect compensability (Ockey & French, 2014). A similar generalization can be made in the current study regarding nonnative raters' nonnative speech ratings from a different L1 background.

If the rater's L2 competence is an important factor influencing the rater, then future research needs to consider which aspect of the term L2 competence is more relevant to foreign accent rating (e.g., proficiency in the L2 or familiarity with accented speech). The current data suggests that perhaps a rater's familiarity with a particular foreign accent is maybe more important than their proficiency in the L2.

5.6.3. Language Pedagogy

The findings from the current study suggest that experience with a specific foreign accent may have facilitated a better understanding of the accented speech, which resulted in more accurate judgments of the speech. The results demonstrated that Arabic and Mandarin raters did not differ significantly from native raters when they were rating speech produced by nonnative speakers from the same L1 background as them (i.e., the nonnative raters). Further, the results showed that nonnative raters were not as accurate as native English raters when rating native English speech. That is, nonnative raters did not always identify native speech to be native.

Previous research has indicated that exposure to variable nonnative and native accents improves nonnative learners' perception and leads to faster adaptation to a wide range of accents, and thus facilitate better processing, comprehension, and intelligibility of these accents (e.g., Bradlow et al., 1997; Bradlow & Bent, 2003, 2008). Previous research has also shown that nonnative raters with repeated exposure to a wide variety of native accents (i.e., regional accents) may better understand what constitutes native speech (e.g., Bradlow et al., 1997; Nishi & Kewley-Port, 2007; Schoonmaker-Gates, 2012;).

These findings have implications for second language pedagogy, particularly listening and pronunciation instruction. For example, if nonnative learners are exposed to various native and nonnative accents, this may help develop their perception in the target language, facilitating better comprehension and intelligibility.

For example, one of the questions Derwing, Rossiter, and Munro (2002) examined relates to the effects of accent familiarity and explicit linguistic training on the intelligibility and comprehension of accented speech. They had two experimental groups of native English speakers who participated in Vietnamese accented English comprehensibility and intelligibility tasks. Participants in the first group were exposed to Vietnamese accented English and received explicit linguistic training on the features of a Vietnamese accent (e.g., consonant clusters, and final consonant deletion). Participants in the second group were exposed to the same Vietnamese accented English files but received no linguistic training on a Vietnamese accent's characteristics. All participants were pre-tested, exposed to the same Vietnamese accented English audio files for eight weeks, and were post-tested. Their results showed that comprehension and intelligibility significantly improved from the pre-test to the post-test for both groups with no differences.

Such a finding suggests that perhaps exposure to accented speech alone could improve comprehension and a more accurate perception of accented speech (i.e., intelligibility). Taken together, exposure to different native and nonnative accents in the target language may be beneficial for the language learners' L2 competence.

5.6.4. Spoken Language Assessment

The current study results revealed that nonnative raters rated nonnative speech from speakers with shared L1 background differently from nonnative speech from a non-shared L1 background, indicating that perhaps familiarity with the L1 or the type of L2 accent affects their rating. It was found that Arabic raters rated Mandarin accented English to have more foreign accent than both English and Mandarin raters. Similarly, Mandarin raters rated Arabic accented English to have more foreign accent than both English and Arabic raters. This finding has important implications for spoken language assessment because if the examiner's familiarity with a specific foreign accent affects his assessment of the test taker, especially if the examiner is a nonnative speaker of the target language. This issue may introduce implicit or explicit biases that could benefit or harm the test taker's chances.

This finding is in line with previous research on language assessment and testing, which has revealed that examiner's familiarity with the test taker's accent influences their speaking skill assessment. For example, Carey, Mannell, and Dunn (2011) have found that IELTS examiners who share an L1 with the test takers tend to assign higher pronunciation scores to those test-takers. This effect is not necessarily limited to a shared L1 background. For example, Winke, Gass, and Myford (2013) found that native English TOEFL raters who studied an L2 (e.g., Spanish), rate native speakers from that L2 (native Spanish speakers) more leniently (i.e., assign higher pronunciation scores). These biases may be due to the phonological adaptation to the test-takers L1 or the test-takers interlanguage.

Taken together, spoken language assessment tests need to take the shared L1 background between the examiner and test taker and the familiarity with a specific accent into account when developing the tests, training examiners, and assigning them to test takers. Further, since the examiner's possible biases are presumably due to their judgment or awareness of phonological features in the test taker's speech, training that highlights these issues, and how to minimize their effects is warranted.

5.7. Conclusion and Future Directions

In this dissertation, an accent rating experiment was designed and carried out to examine whether rater-specific characteristics influence foreign-accented speech rating. It particularly examined whether native and nonnative raters from different L1 backgrounds differ in their evaluation of the degree of foreign accent in English speech produced by native and nonnative speakers from different L1 backgrounds. It also examined whether nonnative raters form different rater groups differ from each other due to their shared L1 status with the speaker. Further, it examined whether trained native raters differ from untrained naïve raters. Finally, it explored the effects of other rater-specific factors such as the nonnative rater's degree of foreign accent, L1 use, and L2 length of residence.

The analysis of the differences between trained native English raters and naïve native English raters revealed that both groups' ratings were strongly correlated, indicating a similar pattern of their ratings. The minimal differences found between the two native groups are not suggestive of different rating behaviors. However, further research is needed in this area. In particular, all the untrained native raters in the current study live in multicultural communities where various foreign accents are spoken, which may have benefited them. Better control needs to be done if such a comparison between the two native rater groups is further explored.

The current study also found that native and nonnative raters differed in their rating of foreign accent. Native raters rated speech from all nonnative speaker groups to have less accent than nonnative raters did. Further, native raters were more accurate than nonnative raters in identifying native speakers. In addition, the study found an interesting pattern in the ratings of the nonnative raters. In particular, when nonnative raters were rating speech from speaker sharing an L1 with them, their ratings were significantly lower than when the same speech was rated by nonnative raters from another L1 background but not significantly different from native raters rating the same speech. This suggests that the shared L1 background status between the nonnative speaker and rater influences their ratings. More nonnative raters from different L1 backgrounds need to be included to confirm these findings and its generalizability to other nonnative accents.

As for the nonnative rater's self-reported L1 use, and own degree of accentedness, the current study found that these factors did not have significant effects on the rating, nor were they found to correlate with the raters' ratings of others' accents. In the future, more rater-specific characteristics need to be included. Previous research has generally found the length of residence and proficiency to affect the rating of foreign accent. Future research needs to consider raters with variable lengths of residence and develop objective measures for proficiency. It is also recommended that future studies include an accent familiarity questionnaire for the raters. Finally, future research could benefit from a qualitative component in the form of raters' feedback or think-aloud protocol detailing the thought process behind the accent judgments. Such a qualitative component may provide significant insights into how native and nonnative raters make their judgments.

Appendix A. Stella Passage

“Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.”

Appendix B. Demographic Information Questionnaire

Participant ID Code:

1. Age: 2. Sex:

3. What is your first language? (please specific)

4. Education (highest degree obtained, or school level attended):

5. Country of origin:

6. Country of residence:

7. Do you speak a second language? If so, what is your second language?

8. At what age did you start to speak your second language?

9. How did you learn your second language?

    ○ Naturalistically

    ○ Academically

10. How long have you lived in an English-speaking country?

11. Write down the name of the language in which you received instruction in school, for each schooling level:

- Primary/Elementary School _______________
- Secondary/Middle School _______________
- High School _______________
- College/University _______________

12. Estimate, in terms of percentages, how often you use or listen to your native language, your second language, and other languages per day in all daily activities combined. Total should equal 100%:

- Native language ___________%
- Second language __________%

Other languages __________%   (Specify _______________)

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4), 390-412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3), 255-278.

Bates, D., Maechler, M., Bolker, B., & Maechler, M. (2013). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999–2.

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America*, 114 (3), 1600-1610.

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

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707-729.

Borrie, S., Mcauliffe, M., Liss, J.,O'Beirne, G., & Anderson, T. (2013). The role of linguistic and indexical information in improved recognition of dysarthric speech. *The Journal of the Acoustical Society of America*. 133. 474-82.

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *University of Hawai'I Second Langauge Studies Paper* 21 (2).

Carey, M., Mannell, R., & Dunn, P. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*. 28. 201-219.

Coulmas, F. (1981). Introduction: The concept of native speaker. *A festschrift for native speaker*, 1-25.

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility: Nonnative listeners' perspectives. *Journal of Second Language Pronunciation*. 2(2), 160-182.

Davies, A. (1991) *The native speaker in applied linguistics*. Edinburgh: Edinburgh University Press.

Davies, A. (2003*). The native speaker: Myth and reality*. Bristol: Multilingual Matters.

Davies, A. (2004). 17 The Native Speaker in Applied Linguistics. *The handbook of applied linguistics*, 431.

Derwing, T. & Munro, M. (1997). Accent, intelligibility, and comprehensibility: Evidence for four L1s. *Studies in Second Language Acquisition*, 20, 1-16.

Derwing, T. & Munro, M. (2009). Comprehensibility as a factor in listener interaction preferences in the workplace. *Canadian Modern Language Review* 66, 181-202.

Derwing, T. & Munro, M. (2001). What speaking rates do non-native listeners prefer? *Applied Linguistics* 22, 324-337.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning,* 63, 163-185.

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam: John Benjamins.

Derwing, T. M., Rossiter, M. J., & Munro, M. J. (2002). Teaching native speakers to listen to foreign-accented speech. *Journal of Multilingual and Multicultural Development*, 23(4), 245-259.

Flege, J. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America*, 76, 692-707.

Flege, J. (1988). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America,* 84, 70-79.

Flege, J., Bohn, O., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25, 437-470.

Flege, J., Frieda, A., & Nozawa, T. (1997). Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics*, 25, 169-18

Flege, J. E. (1999). The relation between L2 production and perception. In *Proceedings of the XIVth International Congress of Phonetics Sciences* ,1273-1276). by J. Ohala, Y. Hasegawa, M. Ohala, D. Granveille & A. Bailey.

Flege, J., Yeni-Komshian, G., & Liu, S. (1999). Age constraints on second language learning. *Journal of Memory and Language*, 41, 78-104.

Fayer, J., & Krazinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, 37, 313-326.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). Package 'irr': Various Coefficients of Interrater Reliability and Agreement. 2012.

Gao, Z. (2019). *Weighing Phonetic Patterns in Non-Native English Speech* (Doctoral dissertation, George Mason University).

Guion, S., Flege, J., Liu, H., & Yeni-Komshian, G. (2000). Age of learning effects on the duration of sentences produced in a second language. *Applied Psycholinguistics*, 21, 205-228.

Hadden, B. L. (1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning*, 41(1), 1-20.

Harding, L. (2012). Accent, listening assessment and the potential for a shared L1 advantage: A DIF perspective. *Language Testing*, 29(2), 163–180.

Hayes-Harb, R., Smith, B. L., Bent, T., & Bradlow, A. R. (2008). The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts. *Journal of Phonetics*, 36, 664–679.

Hsieh, C. N. (2011). Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 9, 47-74

Huang, B., & Jun, S.-A. (2015). Age matters, and so may raters: Rater differences in the assessment of foreign accents. *Studies in Second Language Acquisition*, 37(4), 623–650.

Huang, B., Alegre, A., & Eisenberg, A. (2016). A Cross-Linguistic Investigation of the Effect of Raters' Accent Familiarity on Speaking Assessment. *Language Assessment Quarterly*. 13. 25-41.

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, 10(2), 135–159.

Jiao, D., Watson, W., Gig-Jano Wong, S., Gnevsheva, K. & Nixon, J. S. (2019). Age estimation in foreign-accented speech by non-native speakers of English. *Speech Communication* 106. 118-126.

Julkowska, I., & Cebrian, J. (2015). Effects of listener factors and stimulus proper- ties on the intelligibility, comprehensibility and foreign accentedness of L2 speech. *Journal of Second Language Pronunciation*, 1(2), 211–237

Jun, H., & Li, J. (2010). Factors in raters' perceptions of comprehensibility and accentedness. *Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference*. 53-66.

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly*, 9, 249–269.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2), 155-163

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, 64, 459–489.

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*. 26(2). 187-217.

Kraut, R., & Wulff. S. (2013). Foreign-accented speech perception ratings: A multifactorial case study. *Journal of Multilingual and Multicultural Development*, 34, 249–263.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82(13).

Lenth, R. (2018). Emmeans: Estimated marginal means. *Aka Least-squares Means, R.*

MacKay, I., Flege, J., & Imai, S. (2006). Evaluating the effects of chronological age and sentence duration on degree of perceived foreign accent *Applied Psycholinguistics*, 27, 157-183.

Major, R. (1987). A model for Interlanguage pronunciation. In G. Ioup & S. Weinberger (Eds.), *Interlanguage Phonology: The acquisition of a second language sound system*. Cambridge: Newbury House.

Major, R. (2001). *Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology.* Mahwah, NJ: Erlbaum.

Major, R. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition,* 29, 539-556.

McArthur, T. (1992). *The Oxford Companion to the English Language*. New York: Oxford University Press.

McCullough, E. A. (2013). Perceived foreign accent in three varieties of non- native English. *Ohio State University Working Papers in Linguistics*, 60, 51-66.

Medgyes, P. (1992). Native or non-native: Who's worth more? *ELT Journal*. 46. 340-349

Munro, M. J. (1993). Productions of English vowels by native speakers of Arabic: acoustic measurements and accentedness ratings. *Language and Speech* 36, 39-66.

Munro, M. J. (1998). The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition*, 20(2), 139-154. Chicago

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 45(1), 73–97.

Munro, M. J., & Derwing, T. M. (2001). Modelling perceptions of the comprehensibility and accentedness of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition* 23, 451-468

Munro, M. J. (2003). A primer on accent discrimination in the Canadian context. *TESL Canada Journal* 20 (2), 38-51.

Munro, M. J. (2008). Foreign accent and speech intelligibility. In Hansen Edwards, J. G. & Zampini, M. L. (Eds.). *Phonology and Second Language Acquisition* (pp. 193-218). Amsterdam: John Benjamins.

Munro, M. J., Derwing, T. M., & Morton, S. L.(2006). The mutual intelligibility of foreign accents. *Studies in Second Language Acquisition* 28, 111-131.

Munro, M. J. (2011) Intelligibility: Buzzword or Buzzworthy? In. J. Levis & K. LeVelle (Eds.). *Proceedings of the 2rd Pronunciation in Second Language Learning and Teaching Conference.* Ames, IA: Iowa State University.

Munro, M.J. (2018). Dimensions of pronunciation. In O. Kang, R. Thomson, & J. Murphy. The *Routledge Handbook of Contemporary English Pronunciation*. pp. 413-431. New York: Routledge.

Munson, B., McDonald, E. C., DeBoe, N. L., & White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics*, 34(2), 202-240.

Murry, I. R., and Arnott, J. L. (1993).Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of Acoustical Society of America*. 93:1097-108

Neufeld, G. G. (1980). On the adult's ability to acquire phonology. *Tesol Quarterly*, 285-298

Nishi, K., & Kewley-Port, D. (2007). Training Japanese listeners to perceive American English vowels: Influence of training sets. *Journal of Speech, Language, and Hearing Researc*h, 50(6), 1496.

O'Brien, M. G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning*, 64(4), 715–748.

Ockey, G. J., & French, R. (2014). From one to multiple accents on a test of L2 listening comprehension. *Applied Linguistics*, 37(5), 693-715.

Piske, T., MacKay, I., & Flege, J. (2001). Factors affecting degree of foreign accent in an L2: A Review. *Journal of Phonetics*, 29: 191-215.

Riney, T. & Takagi, N. (1999). Global foreign accent and voice onset time among Japanese EFL speakers. *Language Learning*, 49(2), 275-302.

Riney, T. J., Takada, M., & Ota, M. (2000). Segmentals and global foreign accent: The Japanese flap in EFL. *Tesol Quarterly*, 34(4), 711-737.

Saito, K., Trofimovich, P., & Isaacs, T. (2015). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, 1–25

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, 37(2), 217–240.

Sales, R. (2012). *Perception of foreign accented speech: The roles of familiarity and linguistic training*. PhD dissertation university of North Texas

Schoonmaker-Gates, E. (2012). *Perception of foreign accent in Spanish by native and nonnative listeners: Investigating the role of VOT and speech rate* (Doctoral dissertation, Indiana University).

Scovel, T. (1969). Foreign accents, language acquisition and cerebral dominance. *Language Learning*, 20, 245–253.

Seliger, H. W. (1978). Implications of a multiple critical periods hypothesis for second language learning. *Second language acquisition research: Issues and implications*, 11-19. Chicago

Simonet, M. (2010). Rating accented speech on continua: Nativeness in speech production in highly proficient bilinguals. In M. Ortega-Llebaria (Ed.), *Selected proceedings of the fourth conference on laboratory approaches to spanish phonology* (pp. 37–46). Somerville, MA: Cascadilla Proceedings Project.

Sheorey, R. (1985). Goof gravity in ESL: Native vs. nonnative perceptions. In *19th annual TESOL convention*, New York

Southwood, H., & Flege, J. (1999). The validity and reliability of scaling foreign accent. *Clinical Linguistics & Phonetics*, 13, 335-349.

Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language learning*, 41(2), 177-204

Thomson, R. (2017). Measurement of accentedness, intelligibility, and comprehensibility. *In Assessment in second language pronunciation* (pp. 11-29). Routledge.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, 15(4), 905–916.

Trofimovich, P., Lightbown, P. M., Halter, R. H., & Song, H. (2009). Comprehension-based practice: The development of L2 pronunciation in a listening and reading program. *Studies in Second Language Acquisition*, 31(4), 609-639.

Weber, A., & Pollman, K. (2010). Identifying foreign speakers with an unfamiliar accent or in an unfamiliar language. In New Sounds 2010: *Sixth International Symposium on the Acquisition of Second Language Speech* (pp. 536-541). Poznan, Poland: Adam Mickiewicz University.

Weinberger, S. (2019). *Speech Accent Archive*. George Mason University. http://accent.gmu.edu

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly*, 47, 762–789.

Winke, P., Gass, S, & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral language. *Language Testing*. 30. 231-252.

Wayland, R. (1997) Non-native production of Thai: acoustic measurements and accentedness ratings, *Applied Linguistics*, 18, 345}373.

Yeni-Komshian, G., Flege, J., & Liu, S. (2000). Pronunciation Proficiency in the First and Second Languages of Korean-English Bilinguals. *Bilingualism: Language and Cognition*, 3(2): 131-149.

Yuan, J., Jiang, Y., Song, Z., (2010). Perception of foreign accent in spontaneous L2 English speech, *Proceedings of Speech Prosody 2010*, 1-4. Chicago.

Xie X, Fowler. (2013). Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English. *Journal of Phonetics*.41:369–378. CA

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs?. *Language Testing*, 28(1), 31-50

Biography

Sahar Almohareb received her Bachelor of Arts in English language and literature from Al-Imam Muhammad ibn Saud University, Riyadh, Saudi Arabia, in 2007. She earned her Master of Arts in Linguistics, in 2014, from Florida International University, Miami, FL.