2015-03-01

The Role of Pronunciation in Speaking Test Ratings

Rui Ma
Brigham Young University - Provo

The Role of Pronunciation in Speaking Test Ratings

Rui Ma

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Lynn E. Henrichsen, Chair
Mark W. Tanner
Troy L. Cox

Department of Linguistics and English Language

Brigham Young University

March 2015

ABSTRACT

The Role of Pronunciation in Speaking Test Ratings

Rui Ma
Department of Linguistics and English Language, BYU
Master of Arts

This study explores the weight of pronunciation in a speaking proficiency test at an English as a Second Language (ESL) Intensive English Program (IEP) in America. As an integral part of speaking, beliefs, practices, and research of pronunciation teaching have experienced shifts over the decades (Morley, 1991). Most studies concerning speaking have focused on intelligibility, comprehensibility, and accentedness of speaking, with attempting to address the role of pronunciation in oral communication. However, the degree to which pronunciation is weighed in determining speaking proficiency levels is unclear (Higgs & Clifford, 1982, Kang, 2013). In an effort to contribute to the understanding of this issue, the current study investigates the relationship between pronunciation and speaking proficiency ratings. The speaking proficiency ratings and pronunciation ratings in vowels, consonants, word stress, sentence stress, intonation, and rhythm of 226 speaking samples from English learners were collected at Brigham Young University's (BYU) English Language Center (ELC). The study confirms that suprasegmentals explain more variance than segmentals in English proficiency, and among those suprasegmental features, only the ratings of sentence stress increase incrementally with the proficiency levels without overlapping among proficiency levels.

Keywords: English pronunciation, speaking proficiency, assessment, pronunciation rubric

ACKNOWLEDGEMENTS

I would like to express my gratitude to my family, my professors, my friends, and my students. Their support and encouragement, and the examples they set for me have been the inspiration for me to move forward.

I would like to thank my family. It has been hard for them to have me study in a faraway country, but they tried their best to be there for me. They were always there when I shared my happiness and struggles, and they assured me that things were going well at home.

I thank my thesis committee members and Dr. Hartshorn. They have made themselves available to me, and they cared for me as a person who had potential and not just as one of their students. I appreciate Dr. Henrichsen's time to meet with me regularly, his encouragement, and his insights on pronunciation. I appreciate Dr. Tanner's advice. I appreciate Dr. Cox's guidance and Dr. Hartshorn's advice on statistical analysis.

I thank the Prices and the Fans for their financial support. Their generosity allowed me to focus on my studies.

Finally, I would like to thank my friends and my colleagues. They contributed a lot to the study and to my professional development. Without their help, I would not have been able to conduct this study or write this thesis.

Table of Contents

List of Tables

List of Figures

Introduction

Most English learners receive instruction with the objective of improving their speaking proficiency. This instruction typically includes vocabulary, grammar, formulaic expressions, pronunciation, and many other things. However, the particular contribution of each of these elements to overall speaking ability on test scores is unknown (Kang, 2013). The current study focuses on examining one of these factors, pronunciation, and its weight in determining overall speaking proficiency ratings.

Among all the factors mentioned above, the most salient factor people notice when a second language (L2) speaker opens his or her mouth and begins speaking is pronunciation. In less than one second, nonnative pronunciation can be identified (Flege, 1984). Nonnative pronunciation may have some undesired consequences (Flege, 1995). Listeners may quickly tag the accented English speaker with labels, such as "slow," "needing help," and other descriptions associated with ineffective communicators. Listeners may in turn adjust their word choice and speed, and might modify their responses to avoid embarrassing the other party. With so much going in the listener's head, communication may be jeopardized. Pronunciation is even an essential issue between nonnative speakers (NNSs), for it has the function of building mutual intelligibility and sociocultural identity in the global scenario (Fayer & Krasinski, 1987; Sifakis, & Sougari, 2005). Some learners may despise native-like pronunciation because it alters their identity, while others seek after what they consider as standard pronunciation.

Another complicating factor is that there is no "standard English pronunciation" because English is spoken by many different populations around the globe. In fact, there are more nonnative speakers of English than native speakers around the world, which appears to be causing native speakers to be losing their ownership of English (Crystal, 1997; Jenkins, 2002,

2006). The English language has gradually become the lingua franca of most parts of the world, and its pronunciation varies from region to region. The term lingua franca is used to describe a language which is used as a medium for two people who speak different first languages (L1s) to communicate (Celce-Murcia, 2013). This phenomenon suggests that English does not have a single norm, but rather, people are establishing local norms within their speech communities (Canagarajah, 2014). As a result, intelligible pronunciation may also differ from region to region, with variability in native speaking norms.

Just as there are different varieties of English pronunciation, pronunciation teaching priorities have shifted over the past few decades (Morley, 1991, Levis, 2005). From the 1940s to the 1960s, language teaching was considered as combining linguistic units into meaningful forms. Pronunciation instruction involving imitation and mimicry of a model was a high priority in the Audiolingual method. However, from the late 1960s to the 1980s, questions and suspicions regarding the efficiency and principles of Audiolingual pronunciation teaching emerged. Many teachers abandoned the teaching of pronunciation. From the mid-1980s to the 1990s, an interest in pronunciation was rejuvenated (Morley, 1991). Papers, journal articles, and resource books came out at this time, exploring the effect of different pronunciation teaching methods. Pennington and Richards (1986) reexamined the status of pronunciation and called for an expanded focus on pronunciation in the context of discourse. According to Morley (1991), researchers started raising the importance of pronunciation in communication and created a basic premise that "[i]ntelligible pronunciation is an essential component of communicative competence" (p. 488). These transformations have led to different ways of teaching pronunciation (Levis, 2005), such as the Bowen technique and other more communicative procedures (Bowen, 1972, 1975; Celce-Murcia, 1987; Celce-Murcia, Brinton, & Goodwin, 2010,

pp. 93, 147-148), the color vowel chart (Taylor & Thompson, 2013) and haptic way of teaching (Acton, 2015).

Despite the advent of new, communicative approaches to teaching pronunciation, the role of pronunciation in speaking proficiency tests is still unclear. According to Kang (2013), "no consensus has been reached regarding to what extent different pronunciation features contribute to the overall ratings of speaking assessment" (p. 10). Because many learners have to be assessed on their English proficiency to receive education and achieve promotion, it is worthwhile to investigate if the time and effort they spend on improving pronunciation will result in benefits such as higher scores on speaking proficiency tests and eventually admission into universities or increased opportunities in the workplace. Unfortunately, the role played by pronunciation in overall ratings of English speaking proficiency is not known by either test administrators or test-takers. For most speaking tests, the only resource that educators and learners have for measuring the effect of pronunciation is a speaking rubric, which may or may not even mention pronunciation features.

Rubrics from well-known tests such as the Test of English as a Foreign Language (TOEFL) do not include detailed pronunciation descriptors. This fact "almost ensures that pronunciation will become a stealth factor in ratings and a source of unsystematic variation in the test" (Levis, 2006, p. 245). Taking into account the importance of pronunciation, such an inconsistent and unsystematic approach to evaluating speaking is not satisfactory. Teachers and learners spend various amounts of time on English pronunciation, but whether their time and effort pay off on the testing results or not is questionable. This problem naturally leads to practical questions, such as: How much time and effort should English language learners spend to effectively improve their pronunciation in comparison with other aspects of speaking? What

pronunciation features should they start with? In an effort to answer these questions, one purpose of this study is to discover the extent to which pronunciation contributes to English speaking proficiency ratings. This study also seeks to contribute insight into researchers' and educators' present understanding of pronunciation teaching and assessment.

In order to provide comprehensive background knowledge, the next chapter defines pronunciation-related terms in the current study, briefly recounts previous studies, lays the rationale for the methodology, and proposes the research questions.

Literature Review

In this section of this paper, some background information on speaking and pronunciation assessment will be reviewed with some attention to the methodology, and also a brief summary of what has been found regarding pronunciation and speaking will be given. First, some terms, i.e. intelligibility, comprehensibility, and accentedness, that are commonly used to measure speaking will be introduced. After that, the term *pronunciation* is defined, following which pronunciation-related research will be reviewed. Since the focus of this study is pronunciation in speaking proficiency tests, some guidelines related to rating pronunciation in standardized tests will be analyzed, and after that, the need for using a pronunciation rubric will be suggested.

Terms in Measuring Speaking

The goal of teaching speaking is to facilitate learners' achievement of communicative competence (Celce-Murcia & Olshtain, 2013). Canale and Swain (1980) proposed a theoretical framework of communicative competence, intended to apply to second language teaching and testing. This framework encompasses grammatical competence, sociolinguistic competence, and strategic competence. To illustrate, achieving grammatical accuracy, adapting language use to different contexts, and coping with communication breakdowns, are all parts of communicative competence. In assessing speaking, "intelligibility," "comprehensibility," and "accentedness" often appear as the main evaluation aspects. What complicates things is that the terms intelligibility and comprehensibility are often used interchangeably in daily life, but are given different definitions in research studies.

Intelligibility. Isaacs and Trofimovich (2012) pointed out that Levis (2006) mentioned there were narrow and broad definitions of intelligibility. A narrow definition of intelligibility

was given by Munro and Derwing (1999, p. 289) as "the extent to which a speaker's message is actually understood by a listener." There have been a number of ways to measure intelligibility. It is commonly measured by the accuracy of orthographically transcribed L2 speech by listeners, and sometimes methods such as comprehension questions and true-false statements are used (Isaacs & Trofimovich, 2012; Derwing & Munro, 2009). On the other hand, a broad definition of intelligibility was given by Derwing and Munro (2009, p. 479) as "the degree of a listener's actual comprehension of an utterance." The broad definition was from the listeners' perspectives and was "not usually distinguished from closely related terms such as comprehensibility" (Levis, 2006, p. 252).

Comprehensibility. Different definitions of comprehensibility have been given in different studies. Smith and Nelson, defined comprehensibility as "word/utterance meaning (locutionary force); word/utterance is said to be incomprehensible when the listener can repeat it (i.e., recognizes it) but is unable to understand its meaning in the context in which it appears" (as cited in Gallego, 1990, p. 221). An utterance, therefore, is said to be comprehensible when it conveys the speaker's intention without the speaker pronouncing each sound distinctly as though from the dictionary.

Comprehensibility was also defined by Munro and Derwing (1999) as listeners' perceptions of understanding. In a number of their studies (Derwing & Munro, 1999; Derwing, Munro & Wiebe, 1997; Munro & Derwing, 1995, 1999, 2001), they use a 9-point Likert scale to measure this construct. In another study, Derwing and Munro (2009) defined comprehensibility as "the listener's perception of how easy or difficult it is to understand a given speech sample" (p. 478). Their study showed that the comprehensibility of a speech sample coincided with the amount of time and effort spent deciphering the utterance.

From the discussion above, comprehensibility, can be interpreted as getting meaning across, as opposed to intelligibility which is getting each sound across. In contrast to intelligibility which is usually measured by the accuracy of transcription, comprehensibility is usually measured by the ratings of native speakers (NSs) on a scale. By examining these various studies, it is clear that comprehensibility and intelligibility are two different terms in the research.

Accentedness. Another dimension in speaking is accentedness. Accent has been defined as "the ways in which [an L2 speaker's] speech differs from the local variety of English and the impact of that difference on speakers and listeners" (Derwing & Munro, 2009, p. 476). In light of this statement, everyone speaking English has an accent (Derwing & Munro, 2009; Riney, Takagi, & Inutsuka, 2005). Accent has degrees (Flege, 1995), and a speech sample can be described as being "very accented" or "little accented." Accentedness is usually measured using a Likert scale. Perfectly intelligible (transcribed perfectly) speech may have various degrees of accent as well (Derwing & Munro, 2009).

With the definitions of intelligibility, comprehensibility, and accentedness provided, researchers have explored the relationships among them. Numerous studies have found that intelligibility and comprehensibility correlated with each other, but they had little relation to accentedness (e.g. Derwing & Munro, 1997; Munro & Derwing, 1995, 1999). Even though an unintelligible and incomprehensible utterance can be perceived as heavily accented, a heavily accented speech sample is not necessarily unintelligible or incomprehensible, which serves as evidence to the fact that English interlocutors who grow up speaking different varieties of English are, in general, mutually intelligible. This realization led to a shift in the focus of teaching pronunciation from traditional accent reduction to intelligibility and comprehensibility. Accordingly, it is easy for people to judge pronunciation based on intelligibility and

comprehensibility measures. However, the process of assessing pronunciation through comprehensibility and/or intelligibility involves not only the speaker but also the listener (Levis, 2006). The listener's pronunciation norm, background knowledge, and experience are factors that could influence the rating results. Therefore, the comprehensibility and intelligibility measures of a speaking sample should not be used as measures of pronunciation.

Pronunciation

Pronunciation is an essential term in this research. It is, therefore, necessary to define it. In previous studies, even though pronunciation was noted, the detailed definition in each might differ. In general, pronunciation assessment consists of accuracy in segmentals and suprasegmentals (Goodwin, 2013) and sometimes fluency (Kang, 2010; Kang, Rubin, & Pickering, 2010).

Traditionally, pronunciation is related to expressing referential meaning. Individual sounds with the stress and intonation patterns of the target language form higher level meanings (Pennington & Richards, 1986). Pennington and Richards (1986) specified three types of pronunciation features: segmental features, voice-setting features, and prosodic features. From a micro-perspective, segmental features consist of individual sounds (i.e. vowels and consonants) and other types of features (e.g., aspiration). Voice-setting features "refer to general articulatory characteristics of stretches of speech" (Pennington & Richards, 1986, p. 209). These features are habits people form when they speak. For example, some people tend to round their lips more. People in North America generally create resonance with their lips, while other people may use other organs, such as, the throat, more. Prosodic features involve prosody, or suprasegmentals (i.e. stress and intonation), along with "the related coarticulatory phenomena of the blending and overlapping of sounds in fluent speech" (Pennington & Richards, 1986, p. 210).

J. B. Gilbert (2008) regards two of the three above pronunciation features, namely, segmental features and prosody (which is suprasegmental features), as pronunciation, and gives her definition of prosody which is a "combination of both rhythm and melody" (p. 2). According to J. B. Gilbert, rhythm and melody convey the intentions of the speaker. She also proposes a Prosody Pyramid with thought group being the base of the system, and from that level the pyramid narrows going upward with focus word, stress, and peak, or nucleus.

In the current study, the definition of pronunciation corresponds with J. B. Gilbert (2008)'s view, which includes individual sounds and sound patterning beyond individual sounds. In the remainder of this section, the two dimensions of pronunciation, i.e. segmental and suprasegmentals, will be discussed, and the pronunciation features studied in this research will be summarized at the end.

Segmentals. Segmentals, known as individual sounds, are vowels and consonants. These components of the English language are frequently taught in pronunciation books (Celce-Murcia, Brinton, & Goodwin, 2010; J. B. Gilbert, 2012). Studies of segmentals in second language learning started decades ago. Those studies, however, focused primarily on the differences between the pronunciation of L2 learners and that of native speakers, and the proposal and verification of theories to explain the reasons for the differences.

Flege (1995) and his colleagues developed a speech learning model to account for age-related issues in achieving native-like pronunciation. They proposed that adult English learners had difficulties producing segmentals in a native-like manner when they had to master a vowel system more complex than the one in their native language. Similarly, learners also struggled with unfamiliar consonants. Learners need to have long term memory of all those sounds and train their articulators to be able to pronounce a combination of those sounds in speech. It is

proposed that the greater the difference between English and the learners' L1s, the easier they acquire it (Flege, 1995). For those sounds, learners may perceive them as new sounds so they may develop new habits of producing them. On the other hand, for sounds that are similar to sounds in their L1s, learners may perceive those sounds as the same as those in their L1s. As a result, it may be easy for English learners to compensate certain difficult sounds in English with similar sounds from their L1s. Since the premise that native-like speech is not the most important goal in speaking is established, part of the current study explores, in an ESL setting, the influence of vowels and consonants on the overall proficiency ratings in an English speaking test.

Suprasegmentals. Along with segmentals, suprasegmental features are an integral part of communication (Brazil, Coulthard, and Johns, 1980; Celce-Murcia et al., 2010; McNerney and Mendelsohn, 1992; Morley, 1991; Pennington and Richards, 1986). Some studies have found that suprasegmental features were hard for learners to attend to and to apply. Pennington and Ellis (2000) found that suprasegmentals were not factors that Cantonese speakers naturally attended to and that they had a hard time telling the differences between utterances when the only differences were prosodic features. Also, Bansal (1969) found that apart from the frequency of mistakes in consonants and vowels, mistakes in patterns of word stress, sentence stress, rhythm, and intonation caused differences in intelligibility in Indian English. The following subsections detail the suprasegmental features of pronunciation frequently examined in previous studies, including intonation, word stress, rhythm, sentence stress, and fluency.

Intonation. Intonation was compared to a "road sign" by J. B. Gilbert (2008). From that statement, intonation in English conveys important information. It is defined as the "pattern of pitch and stress in the flow of speech" (Nicolosi, Harryman, & Kresheck, 1989, p. 134). According to Brazil (1997), two identical utterances except for intonation differences convey

different speaker's intentions and listener-speaker convergences which are the shared background knowledge or view by the listener and the speaker.

Intonation can be crucial in educational settings. Pickering (2001) looked at the tone choice in international teaching assistants' (ITAs') communication. Tone choice was defined as "the choice of a sustained rising, falling, or level pitch movement on the tonic syllable in the tone unit" (p. 234), and the manifestation of tone choice is intonation. The researcher recorded six ITAs and six native speakers (NSs) teaching assistants (TAs). All the ITAs were native Mandarin speakers. Extracts of two to four minutes from these TAs' classes were analyzed in terms of falling, rising, and level tones, and the functions of these tones were described. Pickering found that although both groups tended to use a large portion of falling tones, Mandarin-speaking ITAs could not use a variety of intonation features effectively to build TA-student rapport or to effectively convey meaning. ITAs failed to avoid the appearance of overt disagreement, to engage the students, or to indicate the assumption that the listeners already knew the information. ITAs adopted more falling or level tones, which made them appear monotonic and thus less engaging in their speech.

This finding is congruent with Wennerstrom's (1997) study in which she investigated the role of intonation in classrooms and intonation in conversation. Wennerstrom (1997) found that a high tone was used for new material or contrast, and that low pitch was employed for shared information and function words. Wennerstrom (1997) also revealed that intonation carries emotions and signals turn-taking. Wennerstrom (1994) examined speakers of Thai, Japanese, and Spanish reading aloud and speaking freely and found that they used less pitch movement to show contrast in situations where English native speakers would, such as to mark thought groups.

From these studies, it appears that intonation has a crucial role in classroom settings and in communication. Nevertheless, the question remains how much variance in speaking proficiency test ratings could by explained by intonation, especially in an ESL setting where speaking samples are produced by speakers from a variety of language backgrounds and where there is a local English norm. In an attempt to contribute to the answers of this question, intonation is one of the pronunciation features investigated in this study.

Word stress. Word stress is defined by Nicolosi, Harryman, and Kresheck (1989) as the "amount of force or strength of movement in the production of one syllable as compared with another; [it] usually results in the syllable sounding longer and louder than other syllables in the same word" (p. 250). This is used as the definition of word stress in the current study. Zielinski (2008) found that native speakers relied on syllable stress patterns as clues to understand unintelligible words in nonnative speakers' utterances. Zielinski analyzed the transcription and comments made by three native English speakers listening to the nonnative speakers' portions of conversations between the author and a Korean speaker, a Mandarin speaker, and a Vietnamese speaker, respectively. In this study, the context was not accessible to the listeners so that they did not have other resources to guess the unintelligible words. As native speakers rely on word stress to assist understanding, nonnative speakers may utilize word stress as well because nonnative speakers are less likely to utilize context clues (Jenkins, 2002).

Based on these findings, word stress serves an important role in keeping the conversation flow without constant interruptions for clarifying meanings. In speaking tests where responses are recorded, raters cannot ask for clarifications when they perform the rating. Therefore, whether better placing word stress would contribute to higher speaking ratings or not is probed in this study.

Rhythm. A suprasegmental feature closely related to word stress is rhythm. The word "rhythm" is used as a musical feature in daily life. Languages also have rhythmic patterning. Dalton and Hardcastle (1977) defines rhythmic patterning as "a temporal sequencing of similar events" (p. 41), which is adopted as the definition for the current study. They explain that the "similar events" could be recurring patterns of more salient syllables than adjacent ones. It is "established by patterns of stress and rate" (Nicolosi, Harryman, & Kresheck, 1989, p. 230). Adams (1979) believed that "command of rhythm is the key to mastery of the spoken language, and inadequate control of this feature [is] the ultimate barrier to fluency and comprehensibility at all levels of usage" (p. 3). Adams came to the conclusion that factors such as the different rhythmic patterning of their L1s, the way of learning English, and inadequacy of speaking English, were the cause of having nonnative rhythm.

Evidence that English learners have difficulties with the rhythm in the English language is found in research. In some literature, English has been classified as a stress-timed language, where stressed syllables appear at equal intervals. Mochizuki-Sudo and Kiritani (1991) found that Japanese speakers had difficulties telling the durational patterns in English measured by an interstress interval. Therefore, some techniques employed by English speakers to achieve stress-timed patterns, such as reduction, are hard for the Japanese speakers of English to detect. English learners may not only have difficulties identifying English rhythm but also employing it in their speech. Wenk (1985) used the term trailer-timed and leader-timed to describe French and English respectively, claiming that the final syllable in the rhythmic groups in French was much longer. Three elicitation methods, imitative reading, guided re-telling, and sentence-final word echoing, were employed. French learners of English did not perform the target language rhythm,

but there was a transfer stage where they produced language rhythm differing from both French and English.

The studies outlined here used the English produced by native speakers as the standard, but how different performances in producing English rhythm affect the perception of raters in rating speaking proficiency is still unclear. Therefore, there is a need to research how, in a proficiency test, rhythm as part of pronunciation plays a role in the ratings.

Sentence stress. The last pronunciation feature investigated in this study is sentence stress. Sentence stress is also known as primary stress (Hahn, 2004), which is commonly used to draw attention to new or contrastive information. The placement of sentence stress may be hard for English learners to identify. Watanabe (1988) found that Japanese speakers were less successful in pointing out the intonational nucleus in English than native speakers. Also, misplacement of sentence stress may reduce communication efficiency. Hahn recorded three versions of an authentic academic lecture in English by a Korean NS. One version had primary stress correctly placed, the second one had primary stress incorrectly placed, and the third one missed primary stress entirely. Thirty English NSs listened to each lecture. While the main task for the listeners was to listen to the lecture and to understand it, they were asked to perform a secondary task which was to click the computer mouse when they heard a sound in the background. Their reaction time was recorded. After that, they recalled the lecture and took a quiz. The study used the reaction time to measure the difficulty of processing the discourse, and the reflection and the quiz were used to measure comprehension. The results showed that the recording with correct primary stress was the easiest to understand and that listeners' comprehension of the three versions decreased. Hahn drew the conclusion that "correct primary stress in extended nonnative discourse facilitates communication" (p. 215).

Even though these studies only researched learners with Japanese and Korean as their L1s, English learners from other L1 backgrounds may face similar challenges of perceiving and using primary stress in English as well. Therefore, sentence stress, as one pronunciation feature English learners have to acquire, is investigated in this study as to how it contributes to speaking proficiency ratings.

Fluency. As an aspect of speaking, fluency is sometimes included as one of the pronunciation features, but it is not one of the pronunciation features in the current study. Speaking fluency was referred to as "smoothness with which sounds, syllables, words, and phrases are joined together during oral language; lack of hesitations or repetitions in speaking" (Nicolosi, Harryman, & Kresheck, 1989, p. 107). Munro and Derwing (2001) found that there is a curvilinear relationship between speaking rate and English NSs' ratings of accentedness and comprehensibility. They obtained naturally produced samples from nonnative intermediate speakers with various L1 backgrounds and digitally expanded and compressed samples from highly proficient Mandarin speakers of English. They found that English NSs preferred a speaking rate produced by NNSs slightly lower than the normal speaking rate of English NSs because it made the utterance more comprehensible and less accented.

The preferred speaking rate by NNSs, however, is different from NSs preferred speaking rate. Munro and Derwing (2001) had a group of Mandarin speakers and a group of speakers from mixed L1 backgrounds listen to Mandarin-accented narratives with the speaking rate altered by the computer. They found that Mandarin listeners preferred the same rate when they listened to Mandarin accented English as English produced by native speakers. On the other hand, they found that the other group of listeners preferred a speaking rate slightly slower than the natural speaking rate produced by the Mandarin speakers. These two studies indicate that depending on

the listeners' familiarity with the speakers' L1s, different speaking rates may make the speaking sample better understood for different listeners.

Furthermore, Munro and Derwing (2001) pointed out that numerous reasons could explain slower speaking rater in L2, "including production problems due to incompletely developed syntactic and morphological knowledge, slower lexical access, and articulatory difficulties that arise in the production of segments and prosodic patterns that are less well established than native ones" (p. 453). As a result, fluency may be more a component of proficiency than of pronunciation, just as organization, vocabulary, etc. Because of the related issues of fluency, it is not defined as part of pronunciation in the current study.

Summary. As discussed above, pronunciation features, namely vowels, consonants, intonation, word stress, rhythm, and sentence stress are aspects of English that English learners need to acquire in order to achieve communication success. Therefore, this study investigates the influence of these six pronunciation features on speaking proficiency test ratings.

Previous Studies

In an effort to investigate the role of pronunciation in speaking, several studies took the approach of evaluating pronunciation in terms of intelligibility and/or comprehensibility, and other studies were conducted to explore the importance of pronunciation in speaking tests. These studies employed various methods to assess pronunciation and to measure speaking. Following is a summary of methods used in these studies.

The role of pronunciation in communication. Several studies regarded comprehensibility or intelligibility as a measure of communication effectiveness, and these studies usually used linguistic measures to correlate either ratings from novice raters or ratings from trained raters with comprehensibility and intelligibility.

In order to explore the influence of linguistic variables on different levels of comprehensibility (listeners' perceptions of understanding) and to draw the criteria of these variables into rating guidelines, Isaacs and Trofimovich (2012) adapted a mixed-methods approach. Speech samples were recorded from 40 French speakers of English with a variety of proficiency levels telling a story. They were then coded in 19 speech measures such as lexical variables and grammatical variables. Among them, phonology was carefully measured by error ratio. The results were correlated with comprehensibility ratings by 60 native speakers on a 9-point scale. Three experienced ESL teachers were also asked to rate comprehensibility on a 9-point scale and to rate the speech factors they attended to when rating, following which they completed a questionnaire. Isaacs and Trofimovich (2012) found that lexical richness and fluency mattered among low-level learners, that grammatical and discourse-level measure differentiated among high-level learners, and that word stress errors were a salient factor in all levels.

On the one hand, the findings call people's attention to word stress and its influence on speaking comprehensibility, but on the other hand, the findings leave people wondering about the perception of the raters rather than careful linguistic measures and about the influence of other pronunciation features on speaking. At the end of that study, it was discovered that one of the three teacher raters did not interpret "comprehensibility" as the ability to understand, but as intelligibility (word-level clarity). Therefore, it urges the current study to include rater training.

In addition to the exploration of suprasegmental features influence on speaking, researchers have tried to investigate the importance of vowels and consonants. Functional load is a concept to describe the function of differentiating minimal pairs. It has been referred to in the literature since 1955, as noted by Brown (1988), but the definition varies subtly. Functional load

is defined so that segmentals having a high functional load are more likely to impede understanding if they are mispronounced (Brown, 1988). Brown proposed the relative importance of the phonemic Received Pronunciation (RP) contrasts, after analyzing 12 aspects of it: cumulative frequency, probability of occurrence, occurrence and stigmatization in native accents, acoustic similarity, structural distribution of phonemes, lexical sets, number of minimal pairs, number of minimal pairs belonging to the same part of speech, number of inflections of minimal pairs, frequency of members of minimal pairs, number of common contexts in which members of minimal pairs occur, and phonetic similarity. These phonemic contrasts are minimal pairs which may hinder communication if mixed up in speaking.

To test the functional load principle, Munro and Derwing (2006) conducted research where native English speakers rated the comprehensibility and accentedness of sentences containing low or high functional load consonants. The result showed that errors of high functional load consonants led to less comprehensibility and that an increase of the number of these errors correlated with a decrease in comprehensibility. This study narrows down the consonants that were included in the pronunciation rubric used in the current study. Nevertheless, this study had high control over the samples. People do not speak sentence by sentence, but people speak in chunks (A. C. Gilbert, 2014). Munro and Derwing's study was conducted in an unnatural environment, and thus the suggestion of pronunciation instruction should be further analyzed. Therefore, the proposed pronunciation rubric does not entirely rely on the functional load of consonants and vowels, rather it has descriptions and uses the minimal pairs with different functional load as common errors.

The role of pronunciation in speaking proficiency. Besides exploring pronunciation in communication, researchers have attempted to address the issue of what role pronunciation plays

in different speaking proficiency levels. Among the few research studies having been done, a variety of methods were adopted, and the results did not show an agreement.

Higgs and Clifford (1982) used feedback from experienced, trained raters on the importance of five possible speaking subskills (vocabulary, grammar, pronunciation, fluency, and sociolinguistic). They hypothesized a model of each subskill's importance at each proficiency level. In the hypothesis, the importance of pronunciation at the six proficiency levels (from Level 0 to Level 5) was not constant. At Level 1 and Level 5, it was hypothesized that pronunciation had higher importance than at other levels.

In order to test the hypothesis, 50 teachers specializing in a number of languages in the Central Intelligence Agency (CIA) Language School rated the importance of all five subskills, and the results were calculated into percentages. The results for all languages were quite different from the hypothesis and different from the results of the German language alone. Pronunciation played a larger role at the beginning levels than proposed. The importance of German pronunciation dropped in the intermediate level and rebounded, while in the synthesized results of all languages, the importance of pronunciation increased slightly and steadily from each level consecutively. It seemed that at certain proficiency levels, pronunciation might play a larger role, but once another proficiency level was reached, pronunciation may not be as important. On the one hand, how the teachers thought of the importance of pronunciation can be drawn from this study, but on the other hand, without raters listening to speaking samples and making judgments, it is hard to conclude that the importance of the five subskills holds true in actual rating.

While perceptions of the importance of speaking subskills are valuable, researchers continue seeking carefully measured subskills to probe their relationship with speaking

proficiency. Kang (2013) acoustically analyzed speech samples from Cambridge ESOL General English Examinations in an attempt to find out the relative impact of pronunciation features on ratings of nonnative speakers' oral proficiency. The samples represented four proficiency levels in CEFR from B1 to C2. She extracted one-minute monologue speech samples from each of the 120 speaking samples. Stress and intonation were analyzed by Computerized Speech Laboratory (CSL), and fluency by the PRAAT computer program. Segmental features were analyzed by two analysts who reached a inter-rater reliability measure of .81. The result showed that "70% of the variance in Cambridge ESOL four proficiency levels was contributed to by pronunciation errors" (Kang, 2013, p. 12). Among these errors, the following features took up a decreasing proportion of contributing to the overall proficiency: stress and pitch, fluency, segmentals, and tone choice. In this study, samples of one-minute long were analyzed. It is doubtful that the length and the quality of the speaking samples were sufficient enough for the raters to make judgments on pronunciation. In an attempt to better explore the role of each pronunciation feature, in the current study, the raters had access to longer speaking samples, and they could select the speaking samples which best represented the pronunciation quality.

Even though Kang (2013) took a very scientific approach in which speech samples were carefully analyzed in instrumental and auditory ways, the way people perceive speaking performance may differ from phonological analysis. In Douglas (1994)'s study, the conclusion was reached that similar quantitative scores on a semi-direct speaking test represent qualitatively different speaker performances. He used quantitative and qualitative data to compare the two. In the study, six graduate students who were all native speakers of Slovak took the AGSPEAK test. The test consisted of five tasks: answering three unscored warm-up questions, completing ten partial sentences (scored for grammar and comprehensibility), answering questions about a

picture, responding to two open-ended questions, and describing a diagram. The last three tasks were scored for pronunciation, grammar, and comprehensibility. Two trained raters rated all the samples independently on pronunciation, grammar, fluency, and comprehensibility. Their ratings were averaged to produce the final score. Transcriptions of the samples were done by a research assistant. The research discussion focused on grammar, vocabulary, fluency, and content and rhetorical organization. These aspects showed that raters' perception of the performance differed from the performance data. Therefore, the current study took the approach that pronunciation features were measured from the perspective of the raters.

To conclude, a variety of data collection and analysis methods have been used in previous studies. Samples have been elicited through reading aloud, narration, and conversation. They have been acoustically analyzed, rated by novice raters and trained raters, and qualitatively evaluated. Despite the fact that all the studies contributed to providing insights on the issue of pronunciation in speaking, the role of pronunciation in speaking proficiency tests is still unclear.

Pronunciation in Speaking Assessment

As mentioned by Levis (2006) and Isaacs and Trofimovich (2012), most standardized tests recognize the importance of pronunciation, but address it in a vague way. The rating guidelines or rubrics of four standardized tests on the assessment of spoken proficiency will be further probed. In this section, two norm-referenced tests, the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS) test, and two criterion-referenced tests, University of Cambridge English for Speakers of Other Languages (ESOL) Examinations and the American Council of Teaching Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) will be analyzed.

The TOEFL rubric. The TOEFL is a widely-used test for academic, employment, and other proficiency-measurement purposes. Usually, universities and colleges in English-speaking countries require applicants to take the TOEFL and achieve a certain score determined by the individual institution. This test is commonly found in an Internet-based format, meaning the test takers complete each part on a computer, including the speaking part, and then the responses are sent to raters via the Internet. The speaking test contains two major parts: there are two independent speaking tasks and there are four integrated speaking tasks. The independent tasks involve only talking about a topic, while the integrated tasks require reading, listening, or a combination of both.

There are separate rubrics for these two types of tasks, and both of the rubrics are categorized into general description, delivery, language use, and topic development (Educational Testing Service, 2004). The test taker gets a single score for each task, and a sum of all scores is converted as the score for speaking. Pronunciation is categorized under delivery. The rubric uses words such as "minor difficulties with pronunciation or intonation patterns," "listener effort," "intelligible," "awkward intonation," "choppy rhythm" (Educational Testing Service, 2004). There are no definitions for pronunciation or intonation. Because there are no guidelines on how to use the rubric, the proficiency ratings of speaking samples with mixed criteria levels would be subject to the rater. Furthermore, from the pronunciation descriptions, each rater has to establish his or her own standard of "awkward" or "choppy."

The IELTS descriptors. The IELTS is another widely used test in school admissions, often used concurrently with the TOEFL. The speaking part takes on a different format. In this test, individual test takers have an oral interview with an examiner, and the process is recorded. There are generally three parts: the first part focuses on everyday topics, the second part focuses

on the ability to sustain a topic, and the last part concerns the ability to express, analyze, and discuss opinions. The responses are scored based on descriptors of the 9 IELTS bands (British Council, IDP: IELTS Australia, & UCLES, n.d.; also see Figure 1). The examinee receives a score for each of four categories: Fluency and Coherence, Lexical Resource, Grammatical Ratings and Accuracy and Pronunciation. In Band 9, Band 8, Band 6, and Band 4, the test-taker uses "a full range," "a wide range," "a range," and "a limited range of pronunciation features." The descriptions of pronunciation at Band 7, Band 5, and Band 3 are based on the descriptions of the Bands directly above and below. For example, the pronunciation of Band 7 "shows all the positive features of Band 6 and some, but not all, of the positive features of Band 8" (British Council, IDP: IELTS Australia, & UCLES, n.d.). When raters evaluate the speech, the key indicators are the amount of strain caused to the listener, the amount of unintelligible speech, and the noticeability of L1 influence. Again, there are no indications regarding how the individual raters define the range of pronunciation features and the degree of intelligibility and noticeability of L1 influence.

Common European Framework of Reference for Languages (CEFR) guidelines. The Cambridge ESOL testing system is based on the CEFR guidelines which are proficiency guidelines tests may base on, and it offers five proficiency certificates of English, the highest to the lowest being C2, C1, B2, B1, and A2. The descriptions of the speaking part of the certificate levels focus on speech organization and the number of communication situations the test taker can handle. Pronunciation at each level is defined by how easy it is to understand, its intelligibility, and L1 influences. In the C2 description, it says "pronunciation is easily understood and prosodic features are used effectively; many features, including pausing and

BRITISH COUNCIL IELTS AUSTRALIA UNIVERSITY of CAMBRIDGE ESOL Examinations

IELTS Speaking band descriptors (public version)

Band	Fluency and coherence	Lexical resource	Grammatical range and accuracy	Pronunciation
9	• speaks fluently with only rare repetition or self-correction; any hesitation is content-related rather than to find words or grammar • speaks coherently with fully appropriate cohesive features • develops topics fully and appropriately	• uses vocabulary with full flexibility and precision in all topics • uses idiomatic language naturally and accurately	• uses a full range of structures naturally and appropriately • produces consistently accurate structures apart from 'slips' characteristic of native speaker speech	• uses a full range of pronunciation features with precision and subtlety • sustains flexible use of features throughout • is effortless to understand
8	• speaks fluently with only occasional repetition or self-correction; hesitation is usually content-related and only rarely to search for language • develops topics coherently and appropriately	• uses a wide vocabulary resource readily and flexibly to convey precise meaning • uses less common and idiomatic vocabulary skilfully, with occasional inaccuracies • uses paraphrase effectively as required	• uses a wide range of structures flexibly • produces a majority of error-free sentences with only very occasional inappropriacies or basic/non-systematic errors	• uses a wide range of pronunciation features • sustains flexible use of features, with only occasional lapses • is easy to understand throughout; L1 accent has minimal effect on intelligibility
7	• speaks at length without noticeable effort or loss of coherence • may demonstrate language-related hesitation at times, or some repetition and/or self-correction • uses a range of connectives and discourse markers with some flexibility	• uses vocabulary resource flexibly to discuss a variety of topics • uses some less common and idiomatic vocabulary and shows some awareness of style and collocation, with some inappropriate choices • uses paraphrase effectively	• uses a range of complex structures with some flexibility • frequently produces error-free sentences, though some grammatical mistakes persist	• shows all the positive features of Band 6 and some, but not all, of the positive features of Band 8
6	• is willing to speak at length, though may lose coherence at times due to occasional repetition, self-correction or hesitation • uses a range of connectives and discourse markers but not always appropriately	• has a wide enough vocabulary to discuss topics at length and make meaning clear in spite of inappropriacies • generally paraphrases successfully	• uses a mix of simple and complex structures, but with limited flexibility • may make frequent mistakes with complex structures, though these rarely cause comprehension problems	• uses a range of pronunciation features with mixed control • shows some effective use of features but this is not sustained • can generally be understood throughout, though mispronunciation of individual words or sounds reduces clarity at times
5	• usually maintains flow of speech but uses repetition, self-correction and/or slow speech to keep going • may over-use certain connectives and discourse markers • produces simple speech fluently, but more complex communication causes fluency problems	• manages to talk about familiar and unfamiliar topics but uses vocabulary with limited flexibility • attempts to use paraphrase but with mixed success	• produces basic sentence forms with reasonable accuracy • uses a limited range of more complex structures, but these usually contain errors and may cause some comprehension problems	• shows all the positive features of Band 4 and some, but not all, of the positive features of Band 6
4	• cannot respond without noticeable pauses and may speak slowly, with frequent repetition and self-correction • links basic sentences but with repetitious use of simple connectives and some breakdowns in coherence	• is able to talk about familiar topics but can only convey basic meaning on unfamiliar topics and makes frequent errors in word choice • rarely attempts paraphrase	• produces basic sentence forms and some correct simple sentences but subordinate structures are rare • errors are frequent and may lead to misunderstanding	• uses a limited range of pronunciation features • attempts to control features but lapses are frequent • mispronunciations are frequent and cause some difficulty for the listener
3	• speaks with long pauses • has limited ability to link simple sentences • gives only simple responses and is frequently unable to convey basic message	• uses simple vocabulary to convey personal information • has insufficient vocabulary for less familiar topics	• attempts basic sentence forms but with limited success, or relies on apparently memorised utterances • makes numerous errors except in memorised expressions	• shows some of the features of Band 2 and some, but not all, of the positive features of Band 4
2	• pauses lengthily before most words • little communication possible	• only produces isolated words or memorised utterances	• cannot produce basic sentence forms	• speech is often unintelligible
1	• no communication possible • no rateable language			
0	• does not attend			

Figure 1. IELTS Speaking band descriptors.

hesitation, are ‘native-like’” (Council of Europe, n.d., p. 74). The native-likeness is hard to define, and perhaps different listeners interpret it differently. L1 features are mentioned in the remaining 4 levels: At the C1 level, the L1 accent does not affect the clarity of the message; at the B2 level, there may be intrusive L1 features; at the B1 level, L1 features may put a strain on the listener; and at the A2 level, it’s difficult to understand. However, how these pronunciation feature affect communication may differ from rater to rater.

In summary, the CEFR guidelines regarding pronunciation give individual raters the choice to decide how they deal with pronunciation, and the guidelines do not provide much information in finding out the role of pronunciation in speaking.

The ACTFL proficiency guidelines. The ACTFL OPI is a standardized procedure for the global assessment of functional speaking ability. Usually the ACTFL OPI score is used to measure what a person can do with the language, which is based on the ACTFL proficiency guidelines. Even in the rubric of this widely accepted test, the pronunciation descriptions of each level are not consistent. There are 11 proficiency levels in the ACTFL proficiency guidelines, namely, Distinguished, Superior, Advanced High, Advanced Mid, Advanced Low, Intermediate High, Intermediate Mid, Intermediate Low, Novice High, Novice Mid, and Novice Low (American Council on the Teaching of Foreign Languages, 2012). In the highest three levels, the guidelines do not specify pronunciation, except that a nonnative accent does not prohibit the speaker to be at the Distinguished level and that Advanced High speakers use “precise intonation to convey meaning” (American Council on the Teaching of Foreign Languages, 2012, p. 5). For the other levels, the descriptions are distinguished by the degree of native speakers’ understanding, familiarity with nonnative speakers, and the influence of their L1s. For example, through Advanced Mid to Intermediate High, the speakers should be “readily understood”,

"understood", and "generally understood by native speakers unaccustomed to dealing with nonnatives" (American Council on the Teaching of Foreign Languages, 2012, pp. 6-7). At the Intermediate Mid level, "speakers are generally understood by sympathetic interlocutors accustomed to dealing with nonnatives" (American Council on the Teaching of Foreign Languages, 2012, p. 7). Intermediate Low to Novice High level speakers can be "generally understood by sympathetic listeners accustomed to dealing with nonnatives" (American Council on the Teaching of Foreign Languages, 2012, pp. 7-9). At Intermediate Low and Novice High levels, pronunciation is "strongly influenced by their first language" (American Council on the Teaching of Foreign Languages, 2012, pp. 8-9). In the lowest level, that is, Novice Low, pronunciation is the main factor that impedes intelligibility.

Using the degree of native speakers' understanding, familiarity with nonnative speakers, and the influence of examinees' L1s can be problematic. However experienced the testers are, they may not be able to imagine how other people would understand the conversation (Szpyra-Kozlowska, Frankiewicz, Nowacka, & Stadnicka, 2005). Actually, the more experienced the testers are, the more likely they may be able to understand the test taker. Additionally, English learners may have more than one language background, and the non-English pronunciation can come from the speaker's third or fourth language rather than the first.

The Need for a Pronunciation Rubric

In spite of the difficulties of adapting a usable and functioning rubric, using a rubric to evaluate pronunciation has many benefits. First, raters' behavior can be better monitored using a shared rubric than individual questions on a Likert scale; second, the ratings represent meaningful descriptions (Rose, n.d.); and third, the ratings can be analyzed using Rasch modeling.

The first reason for using a rubric in the current study is that the rubric can bring the raters to similar understanding of the rating categories. Raters as human beings have different traits and experience, and thus may award different pronunciation ratings when rating the same speaking sample. In Isaacs and Trofimovich's (2011) study, 30 native English speakers majoring in music and 30 native English speakers not majoring in music rated 40 English speaking samples from nonnative English speakers on comprehensibility, accentedness, and fluency, using Likert scales. They found that musical background made a difference in the ratings, and that music majors assigned lower scores for accentedness, particularly for speakers with low language proficiency.

Familiarity with learners' L1s may result in bias in rating pronunciation. In the situation where there are no pronunciation guidelines, Carey, Mannell, and Dunn (2010) collected data from 99 IELTS raters from five geographical areas and found that the extent of the exposure to the speakers' L1s positively correlated with the pronunciation ratings. It may be understandable that familiarity with the speaker's L1 leads to better understanding, which leads to higher ratings. Similarly with trained raters using rubrics, there may be small but solid bias from familiarity with nonnative speakers' language background (Winke, Gass, & Myford, 2012). Winke, Gass, and Myford (2012) used Many-Facet Rasch Measurement (MFRM) to analyze ratings by 107 trained raters who had language backgrounds of Spanish, Chinese, and Korean, respectively. Each rater rated 82 speaking samples from the TOEFL produced by speakers in the above mentioned three language groups. Winke et al. found that Spanish speaking raters tended to give higher ratings to Spanish speaking people and that Chinese speaking raters tended to give higher ratings to Chinese speakers. Completely avoiding raters' experience and rater-examinee

interaction may not be realistic. The current study attempts to minimize such effects through a pronunciation rubric and rating design.

The second reason for using a rubric is that the ratings have meanings. Rather than numbers on a scale, each rating represents a category that is described in the rubric. Using a rubric, a particular rater may be more likely to give the same ratings to a speaking sample at two different times than using a scale because using a scale depends on perceptions, which can change. Therefore, a rating represents a specific description from the rubric and is interpretable.

Another advantage of using a rubric is the ability to use MFRM could be used in the analysis. Rasch analysis is "a type of statistical analysis developed by the Danish statistician George Rasch in 1960, which is based on probability" (Lee, 2012 p. 280). Some of the reasons why a Rasch analysis strengthens the current study are that different facets are measured on a common ruler, that the functionality of the rubric can be diagnosed, and that the reliability of each facet can be evaluated.

The first reason is that MFRM accounts for score variance caused by irrelevant factors that could systematically influence the observed ratings, called facets (Evans, Hartshorn, Cox, de Jel, 2014). For example, in a pronunciation study, the facets might include pronunciation which is the construct of interest, the severity of the raters, the difficulty of the item, and the criteria being used. With classical test theory, careful consideration must be made to ensure the raw scores students receive have the property of interval data. With MFRM, raw scores are converted to a probability scale composed of equally spaced logits (i.e. log odds ratio) that have the property of interval data, and all the facets are on a common ruler. Thus, MFRM provides insight on the relationship between the examinee and the rater. The MFRM also takes into consideration these facets and produces fair average scores. The fair average scores "compensate for the

severity differences between the raters rating each examinee" (Eckes, 2011, pp. 73-74), and thus, they are regarded as closer to the true scores representing the ability of the examinees.

The second reason is that MFRM allows the rubric categories to be diagnosed for future use. The functionality of the rubric could be addressed by the category measurement statistics. There are some requirements for a well-functioning rubric (Eckes, 2011). The basic requirement of a functioning rubric is that the average measures, which represent the average ability of people placed in each category, increase with the category without dropping. The difference between the observed ratings and the expected ratings, represented by mean-square outfit statistics, should not exceed 2.0. Also, the threshold measures, which indicate that the construct has an equal possibility of being placed into either of the two adjacent categories, should advance continuously with the difference between two measures in the range of 1.4 logits and 5.0 logits.

In addition to how the rubric is used, MFRM calculates how well the facets are separated by the ratings into distinct groups. This information is reflected by the separation reliability and the separation strata in the examinee measurement report (Eckes, 2011). Rather than like other reliability statistics showing how reliably the same, the separation reliability shows how reliably distinct those groups are. The closer the reliability is to 1, the more reliably different the groups are. A reliability close to 0 shows individual differences are small. For example, in a study where student pronunciation is facet of interest, a reliability of .8 shows that 64% ($.8^2$) of the variance in scores can be accounted for by pronunciation and that the examinees could be reliably separated by the instrument and rating procedure. The separation strata statistic is an indicator of the number of groups the facet has. For example, if the rater strata statistic were 2.1, the raters could be differentiated into two groups—a lenient group and a severe group. In this example, in order to compare the examinees to each other, fair average scores would need to be used.

Even though there are advantages of using a rubric to measure pronunciation, currently, there is not a pronunciation rubric that is widely used. Previous studies have not used a pronunciation rubric or considered the English learners' L1s, and acoustic measurement and rating scales have been used to measure pronunciation. Because no pronunciation from previous research could be used, the current study used a pronunciation rubric adapted from a presentation rubric of an advanced ESL pronunciation class at university level.

Research Questions

As reviewed above, intelligibility and comprehensibility may be important aspects in communicative success, but they may not serve as pronunciation measurements. Further, rating guidelines from widely used speaking tests are not informative enough to measuring pronunciation. In addition, previous studies were far from conclusive to answer the question of how much pronunciation weighs in speaking proficiency tests. In an attempt to address this question, the current study proposes to use a pronunciation rubric to assess pronunciation, and in the pronunciation rubric, pronunciation features involve vowels, consonants, intonation, word stress, rhythm, and sentence stress.

The focus of the current study is to better discover the weight pronunciation carries in a speaking proficiency assessment. There are two main research questions in this study:

(1) What role does pronunciation play in determining overall speaking proficiency level?

(2) What aspects of pronunciation influence the overall speaking rating most?

Methodology

In order to answer the two research questions regarding the role of pronunciation and the aspects of overall pronunciation in speaking proficiency ratings, a pronunciation rubric was designed. This section of the paper will introduce the instruments, i.e. the speaking test and the pronunciation rubric, the setting where the current study was conducted, and the procedures of data collection and data analysis. Recall the two research questions: (1) What role does pronunciation play in determining overall speaking proficiency level? (2) What aspects of pronunciation influence the overall speaking rating most?

Instruments

Two main instruments were used in the current study to elicit the speaking samples and to generate the proficiency ratings and the pronunciation feature ratings. This section will describe the proficiency test used to elicit speaking samples as well as the development of the rubric the raters used to rate the pronunciation of each speaking sample.

Speaking test. The speaking samples involved in the current study were obtained through the speaking part of a proficiency test administered at the Brigham Young University's (BYU) English Language Center (ELC). The proficiency test measured the students' proficiency in four language skills, namely speaking, listening, reading and writing, and the results of the test were used to place the students at a level of study in the Intensive English Program (IEP). The students took the speaking test via the computers in the computer lab at the ELC. All the students had taken the same form of the test before and were familiar with the procedure. They were asked to follow the instructions on the computer screen to first confirm their ID and then to proceed to give responses to each prompt. The whole test consisted of 12 prompts, and each prompt was displayed on the screen and an audio file containing the prompt was read by the

computer at the same time. As soon as the computer finished reading the prompt, preparation time was given before the response was recorded. Please see Appendix A for the prompts and time allotted for preparation and speaking. The students then heard a "beep" signaling their responses were being recorded.

The students' responses were rated according to a holistic speaking rubric (see Appendix B) developed by the ELC, which was based on the ACTFL proficiency guidelines to evaluate speaking performance. The rubric consisted of three criteria, namely, Text Type, Content, and Accuracy. The raters were instructed to use the rubric in a manner in which they listened for Text Type first and moved on to the next criterion if they could not give a rating. In this rubric, pronunciation belonged to the larger category of Accuracy, but there was no comprehensive description given. The ratings obtained from the raters were further calculated using Rasch modeling to produce fair average scores for each examinee, and the fair average scores were referred to as proficiency ratings.

Rubric development. In order to carry out the current study, the first step of this research process was to develop a functioning pronunciation rubric so that the rubric would help generate ratings for each pronunciation feature. The pronunciation rubric (see Figure 2) was adapted from one developed by Henrichsen (n.d.a)'s pronunciation rubric whose purpose was to evaluate presentations in a university level advanced pronunciation class. Lynn Henrichsen is a professor in the Linguistics and English Language department of BYU. He has lived and conducted research in a number of countries, and the students he has worked with come from nearly 60 countries (Henrichsen, n.d.b). He is now a teacher trainer in Brigham Young University's MA TESOL program as well as a pronunciation instructor. He has written books, book chapters, and journal articles in the area of pronunciation. Due to his contribution and

specialization, Henrichsen's (n.d.a) was used as the base of the pronunciation rubric used in the current study. Henrichsen's (n.d.a) rubric had 11 criteria (speaking generally, vowels, consonants, intonation, word stress, rhythm, sentence stress, pauses, volume, content, and timing). Among all the criteria, those restricted to evaluate presentations specifically were omitted, i.e. speaking generally, pauses, volume, content, and timing. The criterion, speaking generally, evaluates whether the student could annunciate clearly. The criterion, pauses, evaluates the degree of appropriate use of pausing to improve meaning and presentation skills impact. The criteria, volume and content, were omitted because the students' responses were recorded by tested equipment and because content was not the construct of interest in the current study. The criterion, timing, evaluates if the student follows a presentation time limit which did not apply to the current study.

Only pronunciation related criteria from Henrichsen's (n.d.a) rubric were included in the pronunciation rubric used in the current study, namely, vowels, consonants, intonation, word stress, rhythm, and sentence stress, which were pronunciation features that could influence success in oral communication as reviewed in the previous chapter. Common errors of vowels and consonants were added based on functional load principles by Brown (1988). Functional load is a combination of the frequency of the segmentals and conditional probabilities. Brown investigated minimal pairs. The easiest contrasted segmentals were placed as common errors at the first category. The intonation criterion was changed to be more descriptive (see Figure 2). In addition to the changes in criteria, one rating scale category was added to the four scale categories (with each criterion scored from 1 to 4) in Henrichsen's (n.d.a) rubric, resulting in five rating scale categories in the pronunciation rubric used in the current study. A five-scale category

Category	Vowels	Consonants	Intonation	Word stress	Rhythm	Sentence stress
5	Pronounces vowels correctly all the time.	Pronounces consonants correctly all the time.	Uses rising or falling intonation appropriately all the time. Uses intonation to express a variety of meanings, such as apology, sarcasm, etc.	Places stress on the right syllable of multisyllabic words all the time.	Uses stress-timed rhythm naturally all the time.	Places stress on focus words and other key words all the time.
4	Pronounces vowels correctly most of the time.	Pronounces consonants correctly most of the time.	Uses rising or falling intonation appropriately most of the time but sometimes ineffectively.	Places stress on the right syllable of multisyllabic words most of the time, but misplaces it on a few words.	Uses stress-timed rhythm naturally most of the time.	Places stress on focus words and other key words most of the time.
3	Makes inconsistent vowel errors. Common errors: /iː, ɪ/	Makes inconsistent consonant errors. Common errors: /w, v/ /s, z/	Uses rising or falling intonation appropriately most of the time, but intonation impedes understanding.	Places stress on the right syllable of multisyllabic words most of the time, but misplaces it on certain words.	Uses stress-timed rhythm sometimes and syllable-timed rhythm other times.	Places stress on focus words and other key words sometimes.
2	Pronounces some vowels incorrectly consistently. Common errors: /e, ɪ/, /e, eɪ/ /ɑː, aɪ/	Pronounces some consonants incorrectly consistently. Common errors: /f, h/, /t, d/, /k, g/	Uses intonation appropriately sometimes to express emotion, but uses up-rising intonation for both wh-questions and yes/no questions.	Places stress on the right syllable of multisyllabic words most of the time, but misplaces it on a large number of words.	Rhythm is frequently syllable-timed.	Frequently misplaces stress on focus words and other key words.
1	Vowel errors are frequent. Common errors: /e, æ/, /æ, ʌ/ /æ, ɒ/, /ʌ, ɒ/ /ɔ, əʊ/	Consonant errors are frequent. Common errors: /p, b/, /p, f/ /m, n/, /n, l/ /l, r/	Uses rising or falling intonation inappropriately frequently.	Frequently misplaces stress on multisyllabic words.	Rhythm is not demonstrated.	Sentence stress is rarely identified.

Figure 2. Pronunciation rubric used in the pilot study.

was used because the middle category was needed to describe the pronunciation performance with a balance of strengths and weaknesses (Educational Testing Service, 2006).

Piloting of the pronunciation rubric. To verify the functionality of the adapted pronunciation rubric, a pilot study was carried out. The purpose of piloting the rubric was, on a small scale, to find out if the rubric could effectively separate the examinees and to gather information to modify it for the main study.

Seven raters were recruited to rate ten 90-second long speaking responses from the proficiency test used at the BYU's ELC. The raters were all Teaching English to Speakers of Other Languages (TESOL) MA graduate students enrolled in the Ling 671 Teaching Listening, Speaking, and Pronunciation class at BYU. Six of them had taken a graduate-level course dealing with the sound of language, and the other rater had linguistic knowledge of the English language from undergraduate courses. Among all the raters, five of them were nonnative English speakers whose native languages were Spanish, Portuguese, Ukrainian, Russian, and Mandarin Chinese, respectively. The characteristics of the raters were not considered as bias factors in the pilot study because the focus was on determining the extent to which the pronunciation rubric was able to separate different pronunciation abilities.

The raters carried out a fully crossed rating design. As part of a homework assignment in Ling 671, the raters were asked to rate the pronunciation of a set of carefully chosen speaking samples that had a full range of speaking proficiency levels using a pronunciation rubric. The instructions, the speaking samples, and the pronunciation rubric were sent to all the raters via email. The raters were first asked to familiarize themselves with the rubric. Then they were asked to listen to the speaking samples and to give each sample a score for each criterion. The

raters were given 10 days to complete the ratings on their own, following which the raters filled out a spreadsheet to record their ratings.

Findings. To see how well the pronunciation rubric functioned, a MFRM was adopted to analyze the data. The many-facet Rasch measurement took into account facets that could affect the measurement. These facets were, in this case, the examinees, the raters, and the criteria.

Eckes (2011) identifies the characteristics of a properly functioning rubric. These are introduced in the previous chapter. Acknowledging these features, the category statistics showed that the rubric functioned well in most parts: the average measures went up with ascending category (see Table 1). In addition, all outfit mean-square statistics were less than 2.0. However, **not** all the adjacent thresholds were more than 1.4 logits apart in a monotonically growing trend. In Figure 3, the horizontal axis is the ability of the examinees (from - 4.0 to 4.0), the vertical axis is the level of probability (from 0 to 1), and the numbers in the figure are the ratings. Take an examinee of the lowest ability (- 4.0) for an example, the possibility for being rated as a 1 is very close to 1, and there is a little possibility for it to be rated as a 2. The first two levels were only 0.19 logit apart. Figure 3 gives a visual distribution of the levels. Note, however, that an examinee with an ability of -1 has an similar possibility of being rated in Category, 1, 2, and 3.Due to the fact that there were a limited number of samples and raters involved in the pilot study, five distinct category levels in the rubric were kept. The reason was that the rating data could be recoded after they were collected, but prematurely reducing the number of the category levels may decrease the reliability. With the rater materials and the rating procedure used, the separation strata statistic in the examinee measurement report was 6.13 with a reliability of .95. Therefore, the pronunciation rubric and the procedure were considered appropriate to be used in the main study.

Table 1

Category Statistics from the Pilot Study

Category	Absolute Frequency	Relative Frequency	Average Measure	Outfit	Threshold	SE
1	72	18%	-1.71	0.7	NA	NA
2	63	16%	-0.66	1.0	-0.97	0.16
3	131	33%	0.03	0.9	-1.16	0.13
4	104	26%	0.24	1.2	0.34	0.13
5	32	8%	0.08	1.0	1.79	0.20

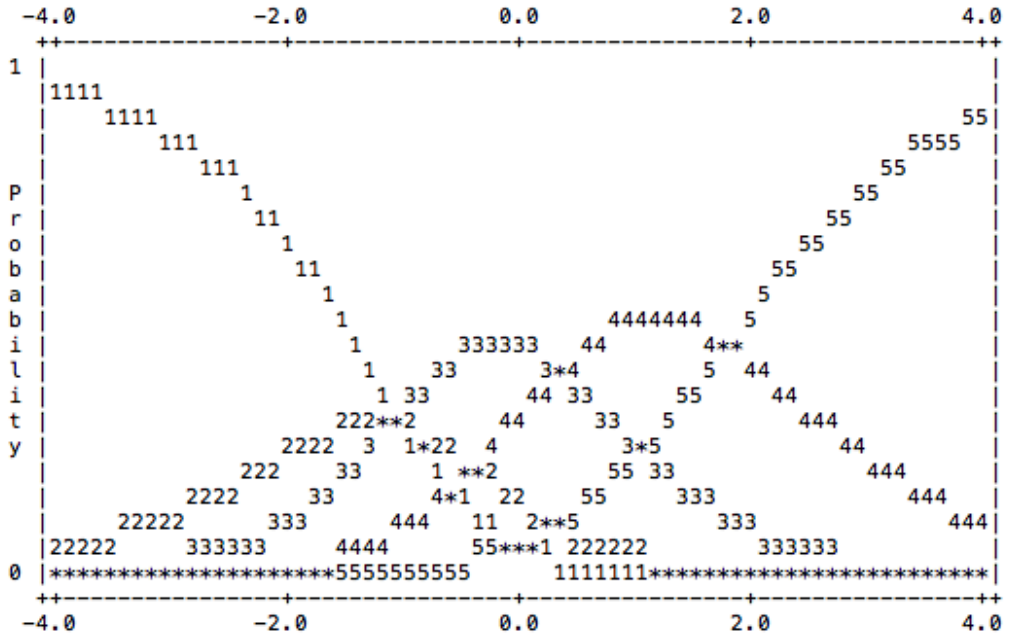

Figure 3. Probability curves of ratings for a given pronunciation ability.

Some raters voluntarily gave feedback on the rubric. One rater said that she was not certain what rhythm meant, and she hoped the rubric would be more explanatory. Another rater commented on the rubric file. She could not see the common errors listed in the rubric because the International Phonetic Alphabet (IPA) symbols were not viewable on her computer. A PDF

file would have worked better. One rater commented that the descriptors of Category 5 were actually describing pronunciation produced by native speakers.

Refining the rubric. In response to the raters' comments from the pilot study, several changes in the descriptors were made (see Figure 4), to make the rubric more descriptive and to better reflect the pronunciation ability of the English learners. The main change was that each criterion became the subject of each descriptor sentence. Also, the descriptors in Category 5 were changed so that the examinee did not have to have perfect native-like pronunciation to achieve this score.

Participating Program, Students, and Raters

The study was conducted at BYU's ELC, an IEP, in Provo, Utah, USA. Eight proficiency levels which corresponded approximately with the ACTFL proficiency levels from Novice Low to Advanced High were taught at the ELC. Among the eight levels, Level 1 and Level 5 were preparation levels for students who lacked proficiency in one or two skill areas to begin the higher level of study. At each placement level, the students took four skilled-based classes a day for four days a week. The instructional program at the ELC was designed in a way that, after one semester of study, students were supposed to be ready for the next level. At the end of each semester (or at the beginning of the semester for new students), the students took a test designed to measure English proficiency in general rather than achievement, and then they were assigned a level for the next semester. The speaking portion of this particular test was used in the current study to determine the speaking proficiency of the students.

In the semester the study was carried out, 238 students were enrolled at BYU's ELC. There were 143 female students and 95 male students with an age range of 17 to 49. The distribution of the students at the beginning of the semester is shown in Table 2.

Category	Vowels	Consonants	Intonation	Word stress	Rhythm	Sentence stress
5	Mispronounced vowels are rare and cause no distraction or miscommunication.	Mispronounced consonants are rare and cause no distraction or miscommunication.	A variety of intonation patterns effectively reflect the speakers' intent (e.g., questioning, apology, sarcasm, etc.)	Misplaced word stress is rare and causes no distraction or miscommunication.	Stress-timed rhythm is used naturally and consistently.	Sentence stress is almost always placed appropriately based on the speaker's communicative intent.
4	Vowel errors occur occasionally (especially in vowel dense contexts) but do not lead to miscommunication.	Most consonants are pronounced correctly most of the time, but troubles with consonant clusters, word-final consonants, etc. persist.	Intonation is employed effectively to express emotion, but one particular pattern is overused.	Misplaced word stress is rare and it only occurs in multisyllabic words.	Stress-timed rhythm is employed naturally most of the time.	Sentence stress is placed correctly most of the time, but sometimes misplaced.
3	Vowel errors (such as /iː, ɪ/) occur frequently and inconsistently but do not usually cause miscommunication.	Frequent but inconsistent consonant errors occur, such as /w, v/, /s, z/	Intonation is usually correct but occasionally misleads listeners.	Misplacement happens in a variety of words, but meaning is not hindered.	Stress-timed rhythm is employed sometimes appears but only unnaturally and with effort.	Sentence stress is employed, but not always correctly (e.g., function words receive stress inappropriately).
2	Some vowels (such as /i, ɪ/, /e, eɪ/, /ɑː, a,/) are consistently confused or mispronounced and cause miscommunication or distraction.	Some consonants (such as /f, h/, /t, d/, /k, g/) are consistently confused or mispronounced and cause miscommunication or distraction.	Rising and falling intonation patterns are sometimes used appropriately but often impede understanding.	Due to frequent and confusing word stress errors, context is greatly needed for the listener to understand the intended meaning.	Rhythm is heavily syllable-timed, but occasionally demonstrates stress-timing.	Sentence stress is rarely used or is frequently misplaced, leading to miscommunication or confusion.
1	Vowel errors (such as /ɛ, æ/, /ɑ, ʌ/, /u, ʊ/, /ɔ, oʊ/) are frequent and distracting and often cause miscommunication.	Consonant errors (such as /p, b/, /p, f/, /m, n/ /n, l/, /l, r/) are frequent and distracting and cause miscommunication.	Intonation is used inappropriately and interferes with communication or is distracting.	Frequent word-stress misplacement causes miscommunication and annoys listeners.	Rhythm is predominantly and strongly syllable-timed (i.e., very "choppy").	Sentence stress is not used to indicate key words in thought groups.

Figure 4. Revised pronunciation rubric used in the main study.

Table 2

Distribution of the Students in the Current Study

ELC Level	ACTFL Level	Number of Students
1	Novice Low and Novice Mid	7
2	Novice High	16
3	Intermediate Low	30
4	Intermediate Mid	64
5	Intermediate Mid	31
6	Intermediate High	62
7	Advanced Low	17
8	Advanced Mid and higher	16

The raters for the speaking portion of the test were 27 teachers at the ELC. Each teacher had some rating assignments at the end of the semester. The number of assignments depended on the number of classes the teacher taught in that semester. The majority of the raters had been rating the speaking portion for many semesters.

The 11 pronunciation raters of this study were a subset of the ELC raters who performed the proficiency rating. The pronunciation raters were volunteers who were interested in the usage of the pronunciation rubric and the results of this study. In addition to giving a holistic proficiency rating, they volunteered to rate the pronunciation of the assigned speaking samples as part of their rating assignments. Among all the pronunciation raters, three were nonnative speakers. The raters' self-reported language proficiencies are summarized in Table 3. All the speaking samples from a total of 238 students were assigned to the volunteer raters. It is important to note that neither the speakers nor the raters were the factors under investigation. The speaking samples and the ratings were elicited to analyze the weight of perceived pronunciation in speaking proficiency tests.

Table 3

Volunteer Raters' Self-Report Proficiencies

Raters	Native	Advanced	Intermediate	Novice
1	English		Mongolian Korean	
2	Spanish	English		
3	Spanish	English		
4	English		German Spanish	Arabic French
5	English			Spanish
6	English	Japanese		
7	Mandarin	English		
8	English	Spanish Portuguese	French	
9	English		Spanish	
10	English		Spanish	French Russian
11	English		Spanish	Korean Portuguese

The proficiency rating design was such that seven of the speaking samples were randomly selected to be rated by all 27 raters and that all samples from Level 4 and Level 5 students were rated by three raters and the other samples were rated at least by two raters. In the process of rating, raters' familiarity of the examinees or examinees' L1s may lead to biased ratings (Carey, Mannell, & Dunn, 2010, Winke, Gass, & Myford, 2012), as is introduced in the previous chapter. In order to minimize rater-examinee interactions, the 11 volunteer raters were assigned speaking samples from learners who were not taught by them and from learners who did not speak their foreign languages, except Spanish. Spanish was the most common foreign language of students at the ELC. Spanish speakers, both students and teachers, were the majority, so it was not possible to avoid Spanish-speaking raters rating samples from learners whose L1 was also Spanish. However, the proportion of the number of samples from Spanish speakers to the total number of rating samples for an individual rater who spoke Spanish was much lower than the proportion of the Spanish speakers to the total enrollment at the ELC which was 50.84%. Table 4 shows the percentage of the number of the samples from Spanish speakers in the total number of rating samples for the 11 volunteer raters.

Table 4

Percentages of Samples with Spanish L1 in the Total Number of Rating Assignments

Raters	Percentages
1	72.00%
2	24.44%
3	24.44%
4	40.00%
5	48.00%
6	71.11%
7	71.11%
8	28.00%
9	40.00%
10	40.00%
11	40.00%

Training Meeting

To avoid any affective bias, a training meeting was held. Training meetings can reduce the number of errors that are caused by familiarity with the students' language background and the raters' experience. Some common rater errors are restriction-of-range effect, halo effect, rater-examinee interaction, severity error, generosity error, etc. (Eckes, 2011; Myford, C. M. & Wolfe, E. W., 2003). The purpose of a training meeting was to let the raters understand and avoid the possible rater errors, know the importance of rating consistency, and build a common understanding of the rubric.

The training meeting was conducted after the calibration which was an ELC procedure and an hour before the actual rating. In the ELC calibration process, all the raters rated 11 proficiency samples from previous semesters. The volunteer raters were also asked to rate the pronunciation of the calibration samples based on the pronunciation rubric. Two of the volunteer raters (Rater 2 and Rater 8) were unable to complete the calibration rating for pronunciation but had the same information packet as was used in the pilot study. The analysis evaluating rubric usage indicated that these two raters behaved within acceptable range. These two raters were

able to participate in the training meeting for general speaking proficiency. In the meeting, each rater received a folder containing paper copies of the assignments and the rubrics (speaking proficiency rubric and pronunciation rubric), handouts containing information about rater errors, and descriptive statistics of the calibration, to discuss different rater errors. After that, the raters understood that they should be consistent. In terms of pronunciation rating, the raters could not be assumed to have the same understanding of the terms (Isaacs & Trofimovich, 2012). Therefore, each volunteer rater received additional handouts with definitions of terms used in the pronunciation rubric and descriptive statistics to inform them of their rating error tendency (see Appendix C). Individuals compared their ratings with the averages to gain a broad picture of their rating error tendency. They were also warned that the descriptive statistics may not have been an accurate reflection of their error tendency.

Procedures

All raters received assignment sheets with the examinees' ID number and rating scales on them, and the volunteer raters received additional copies to record their pronunciation ratings. The speaking samples were stored on computers. Raters had their own account set up in order to listen to the speaking samples assigned. Each speaking sample contained all the responses to the questions on the speaking test from one examinee. Each response was a separate file, and the rater had to click on each file to listen to it. Raters could see the corresponding prompt on the screen while listing to a specific file. The raters were instructed to first give a holistic score based on the proficiency rubric and then give a score for each criterion on the pronunciation rubric. All the raters proceeded with this task at their own pace. To rate proficiency, they were instructed to use the rubric from the left column to the right (see Appendix B), and move to the next column only when the criterion or criteria was not sufficient for them to make their rating. The raters

could choose responses to any questions of certain difficulty they needed to place the rating. After that, the rater would rate the pronunciation according to a response that they deemed best represented the pronunciation of the speaker. Then raters turned in their assignment sheets marked with the ratings, and they also entered their ratings into the computer.

Data Analysis

The actual number of students who took the speaking test was 226. To answer the research questions, two steps of data analysis were followed. In the first step, the functionality of the scale and the reliability of human rating were verified by a Rasch analysis, and the fair average scores of each speaking sample were calculated.

In the second step, regression analyses were carried out to answer the two research questions. To answer the first research question, a simple regression analysis was implemented to see how well the pronunciation fair average scores were able to predict the fair average scores from proficiency ratings. To answer the second research question, a multiple regression analysis was utilized with ratings from each pronunciation feature being the independent variables and the proficiency ratings being the dependent variable. These analyses provided information about how well speaking proficiency can be predicted by pronunciation and the six features investigated in this study, respectively.

Results and Discussion

The current study explored the influence of pronunciation on speaking proficiency ratings. The purpose was to find out (1) how much of the variance in speaking proficiency ratings is accounted for by variance in overall pronunciation and (2) which pronunciation features (i.e. vowels, consonants, intonation, word stress, rhythm, and sentence stress) explain the most variance in the speaking proficiency ratings.

This section presents results from the study in order to answer the research questions. To accomplish this objective, a preliminary Rasch analysis was performed to determine the functionality and the reliability of the pronunciation rubric used by the raters in this study. The findings will be reported for the two research questions: (1) What role does pronunciation play in determining overall speaking proficiency level? (2) What aspects of pronunciation influence the overall speaking rating most?

Research Question 1

The first research question was: What role does pronunciation play in determining overall speaking proficiency level? The functionality and reliability of the rubric as a whole are first presented, and then the research question is answered through analysis.

Phase 1: Functionality and reliability of the rubric as a whole. The five-category rubric (with pronunciation levels scored from 1 to 5) functioned within acceptable parameters for the current study. The outfit mean-square statistics which indicate the difference between the average measure and the expected measure at each rating category did not exceed 2.0 (see Table 5). The average measures at each category and the threshold estimates increased continuously with the threshold estimates of adjacent categories between 1.4 logits and 5.0 logits apart (see

Table 5). The spacing of the thresholds was regularly spaced (see Figure 5). According to Eckes (2011), these category statistics are evidence that the rubric functioned well.

Table 5

Pronunciation Rubric Rating Scale Category Statistics

Category	Absolute Frequency	Relative Frequency	Average Measure	Outfit	Threshold	SE
1	33	1%	-2.72	1.0	NA	NA
2	269	13%	-1.08	1.0	-4.19	0.21
3	720	35%	0.59	1.0	-1.24	0.08
4	726	35%	2.26	1.0	1.40	0.06
5	389	16%	4.19	0.9	4.02	0.08

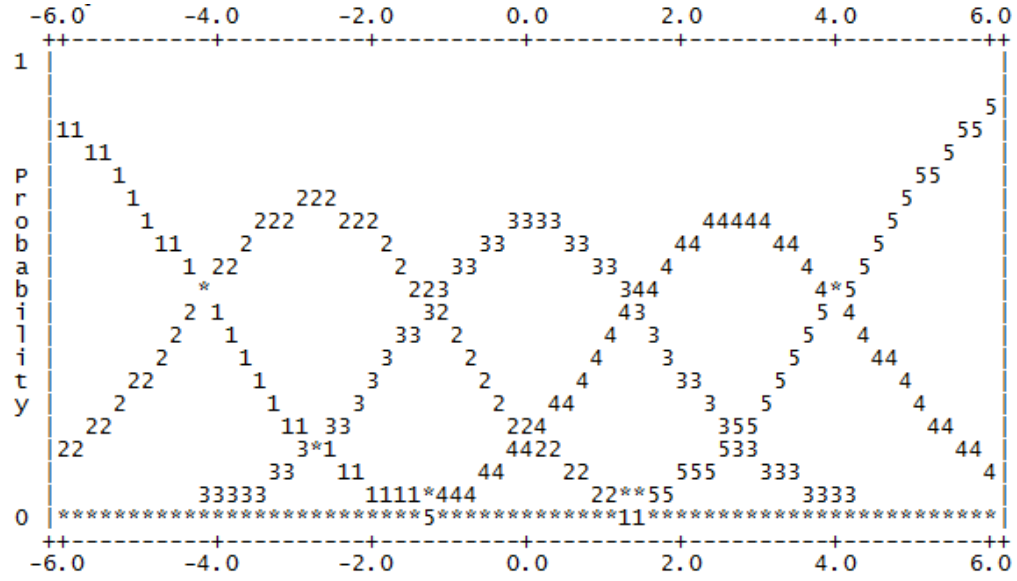

Figure 5. Probability curves for a given pronunciation ability in the main study.

Figure 6 is the variable map that puts all facets, i.e. raters, examinees, and criteria on a single scale. The first column on the left is the scale, serving as a reference for all facets. The second column displays the estimated pronunciation ability of examinees. Each star represents two examinee with higher position indicating higher ability. The third column is the leniency of the raters, and the numbers are the rater IDs. For example, Rater 5 gave the highest ratings which indicative of generosity error, and Rater 4 at the bottom of the scale might have severity error.

The fourth column is the difficulty of the criteria. From the easiest to the most difficult, were word stress, vowels, consonants, sentence stress, intonation, and rhythm. The last column shows the ratings generated by the raters. The dashed horizontal lines are points where there is same possibility of getting rated as either one of the rating above and the rating below the line. An examinee with an average ability (indicated by 0 in the first column) would have less than 50% chance of being rated as 3 by Rater 4 and have more chance of being rated higher than 3 by other raters.

According to the rater measurement report, the 11 raters could be divided into nine distinct groups with a reliability of .99, which means the raters could be reliably grouped into nine groups according to their severity. For raters to be used interchangeable, the reliability would have to have been close to 0. However, the fair average scores can be used to counter the differences in raters, as long as the raters are self-consistent, any rater could generate similar ratings for the same sample. Rater fit statistics between 0.5 and 1.5 are considered "useful fit" (Eckes, 2011). Higher fit statistics show that the rater has more variation than expected, which is called misfit. On the other hand, lower fit statistics show that the rater is predictable and does not provide useful information, which is called overfit. According to Eckes (2011), "misfit is more problematic than overfit" (p.58). In generating pronunciation ratings, Rater 3 was slightly overfit with an outfit mean-square statistic of .43

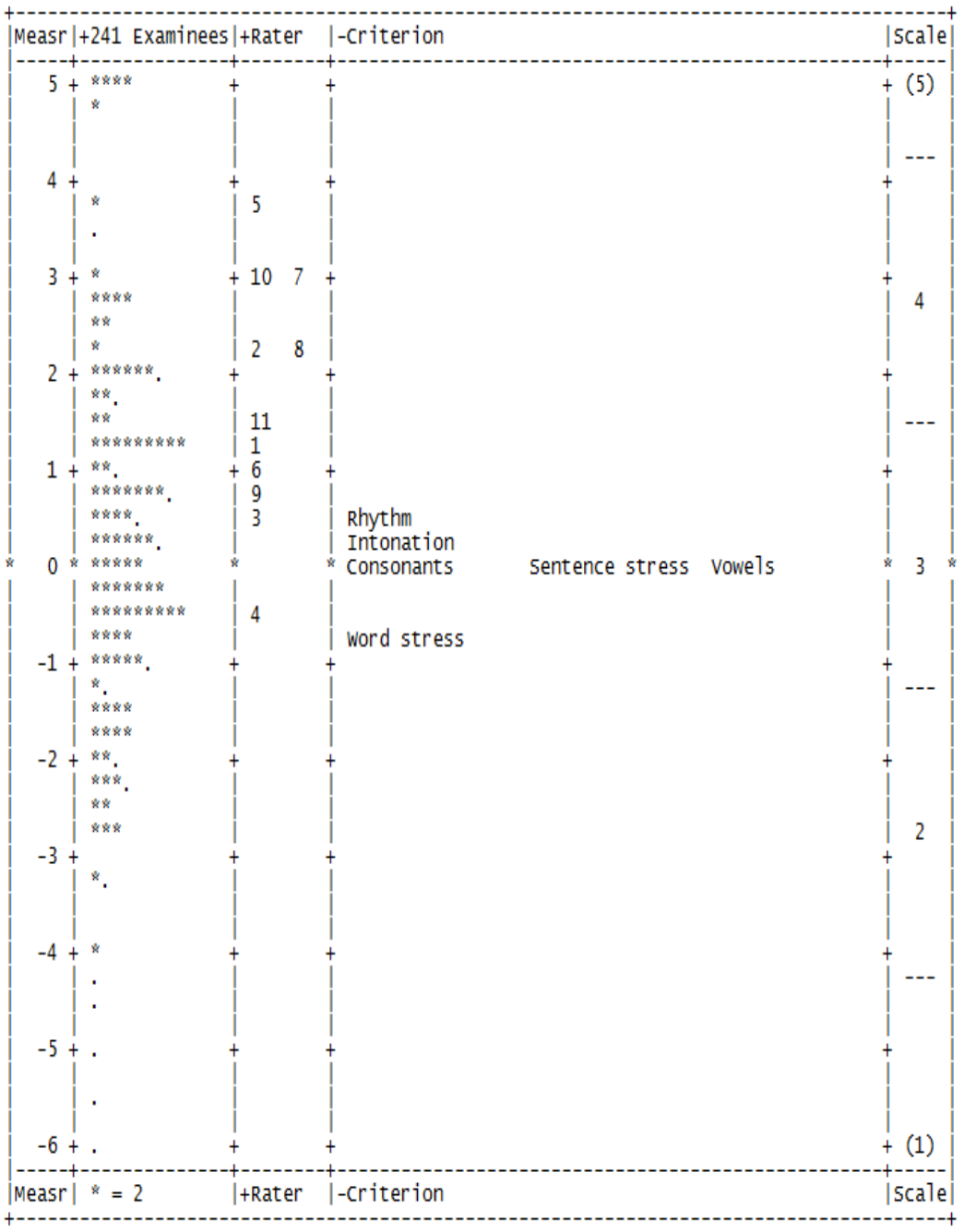

Figure 6. Pronunciation rubric level vertical scale.

Even though the raters did not behave in unison, the examinees could be separated into nearly four different pronunciation groups indicated by the separation strata of 3.68 with a reliability of .86. The results showed that the raters were able to use the rubric well in separating the examinees into different ability levels.

Phase 2: Statistical analysis of pronunciation and proficiency ratings. To answer the question of what role pronunciation plays in determining overall speaking proficiency level, a simple regression analysis was utilized with the proficiency rating of the speaking test being the dependent variable and the pronunciation rating being the independent variable. The positive standardized coefficients beta size shows that there was a positive relationship between pronunciation and proficiency, and the value being .644 indicates that the relationship was relatively strong (see Table 6). From Table 7, this model produced an R^2 of .414 (adjusted R^2 = .411), indicating that 41% of the proficiency rating variance can be explained by pronunciation.

These findings suggested that pronunciation has a strong contribution to determining speaking proficiency level despite the fact that it is less mentioned in proficiency rating than other speaking features, such as organization and content. Because those factors were not included in this study, the extent to which they may co-vary with pronunciation is unknown. Regardless of how other factors might influence speaking proficiency, the current study provides evidence that pronunciation accounts for some of the variance in overall speaking proficiency.

Table 6

Coefficients of Regression Model of Pronunciation Related to Speaking Proficiency

		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	.496	.282		1.760	.080		
	Pronunciation	.953	.076	.644	12.498	.000	1.000	1.000

a. Dependent Variable: Proficiency

Table 7

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df 1	df2	Sig. F Change
1	.644[a]	.414	.411	.78497	.414	156.212	1	221	.000

a. Predictors: (Constant), Pronunciation

b. Dependent Variable: Proficiency

Research Question 2

The second research question was as mentioned: What aspects of pronunciation influence the overall speaking rating most? The usefulness of the categories and the separation reliability at each criterion level are presented first, followed by the findings.

Phase 1: Functionality and reliability of the rubric at criterion levels. In the first phase of answering the second research question, how well the pronunciation rubric was analyzed. Recall the pronunciation rubric (see Figure 4) was composed of six pronunciation features as criteria: vowels, consonants, intonation, word stress, rhythm, and sentence stress. As reported in the first question, the person separation of the rubric as a whole was .86 and each criterion contributes to that overall reliability. The use of the pronunciation rubric will be reported in order of magnitude of differentiation between examinees.

Word stress. For this section, there will be a diagnosis of the rating scale, an analysis of how the raters used the scale, and most importantly to the research question, how well the examinees were differentiated from each other.

The rubric for word stress did not function as expected. Only four out of five rating categories were used by the raters, and Category 1 was seldom used. The average measures advanced continuously, and the outfit mean-square statistics were less than 2.0 (see Table 8). However, the threshold measures were more than 5 logits apart indicating overuse of those categories. From Figure 7, ratings gathered toward the middle categories. This means that either Category 3 or Category 4 or both categories could be split, resulting in more than 5 or 6 categories. For future use of this rubric, the criterion at the word stress level might need to be better defined for raters use it effectively.

Table 8

Criterion Word Stress Rating Scale Category Statistics

Category	Absolute Frequency	Relative Frequency	Average Measure	Outfit	Threshold	SE
1	2					
2	32	5%	-7.51	0.0	NA	NA
3	114	41%	-2.99	0.0	-7.70	0.37
4	109	39%	3.52	0.0	0.45	0.25
5	98	15%	10.28	0.0	7.25	0.32

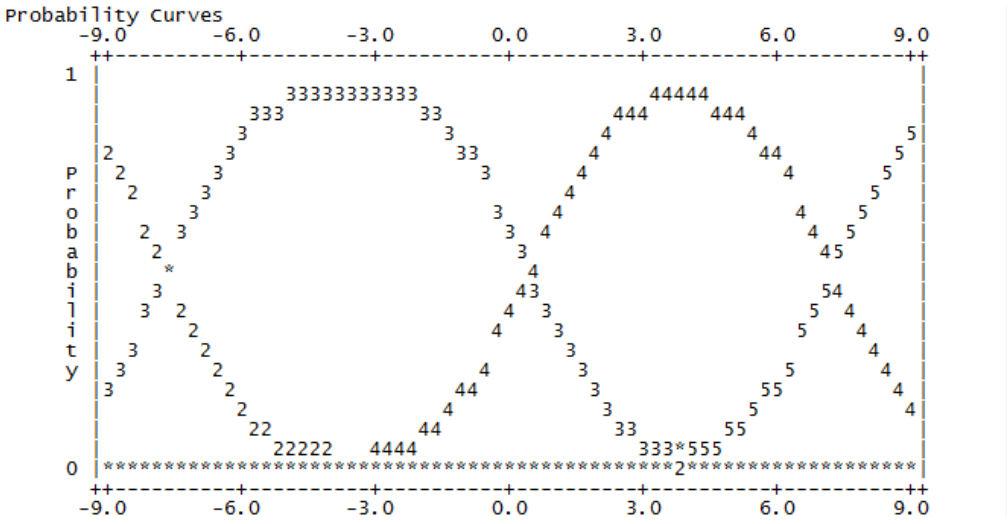

Figure 7. Probability curves of a given ability to produce word stress.

To evaluate how the raters used the rubric, the rater facet was analyzed. All the raters except Rater 1 and Rater 7 used the criterion consistently internally because the fit statistics were between .5 and 1.5. The separation reliability of .98, showing all the raters had different severity in rating word stress (see Figure 8), however since the fit statistics were within the accepted parameters, the fair average can be used to evaluate examinees.

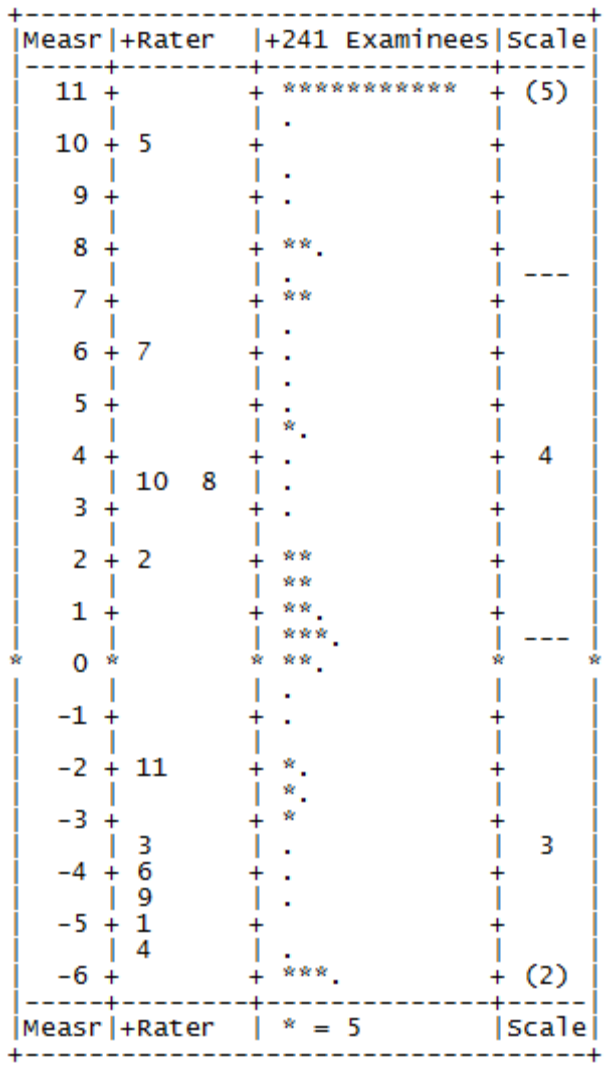

Figure 8. Criterion word stress level vertical scale.

Word stress, as a single criterion, did not separate the examinees very well, as indicted by a separation reliability of .41 or 16.8% of the score variance. This could be an indication that this

rubric criterion could need revisions, the raters might need more training, or it is a skill that examinees acquire early and thus perform more uniformly after a semester of instruction. As a single criterion, word stress was not a strong indicator of different pronunciation ability.

Vowels. The usage of the rubric at the criterion level of vowels resembles that of word stress. Four out of five rating categories were used effectively by the raters. Category 1 was not effectively used because even low ability did not result in getting rated as 1 (see Figure 9). For the four effectively used categories, the mean-square outfit statistics did not exceed 2.0 (see Table 9). The average measures at each category and the threshold estimates increased continuously with the threshold estimates of adjacent categories between 1.4 logits and 5.0 logits apart (see Table 9 and Figure 9). These statistics were within an acceptable range (Eckes, 2011).

Table 9

Criterion Vowels Rating Scale Category Statistics

Category	Absolute Frequency	Relative Frequency	Average Measure	Outfit	Threshold	SE
1	1					
2	47	9%	-4.37	0.9	NA	NA
3	125	42%	-1.74	0.8	-4.42	0.27
4	113	38%	1.85	1.1	0.08	0.19
5	72	11%	4.45	1.0	4.34	0.26

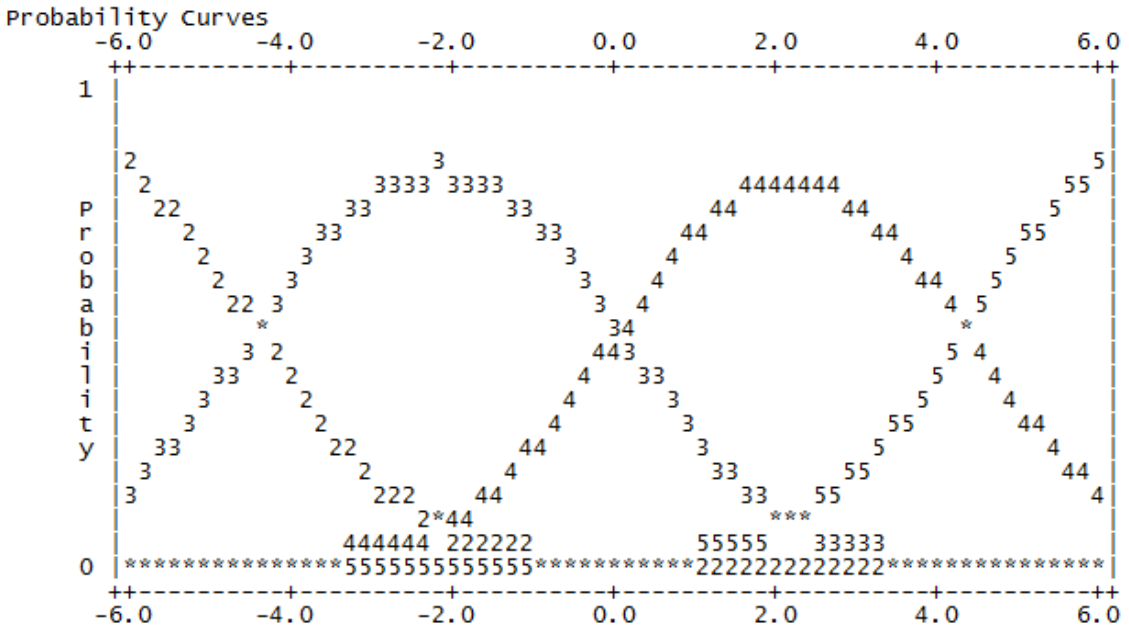

Figure 9. Probability curves of a given ability to produce vowels.

The analysis of the raters attained a separation reliability of .97, indicating that the raters had varying degrees of severity when performing the rating. From Figure 10, Rater 10 was the most generous and Rater 4 was the most severe. Despite differences in severity, the fit statistics suggest that individual raters were consistent within themselves.

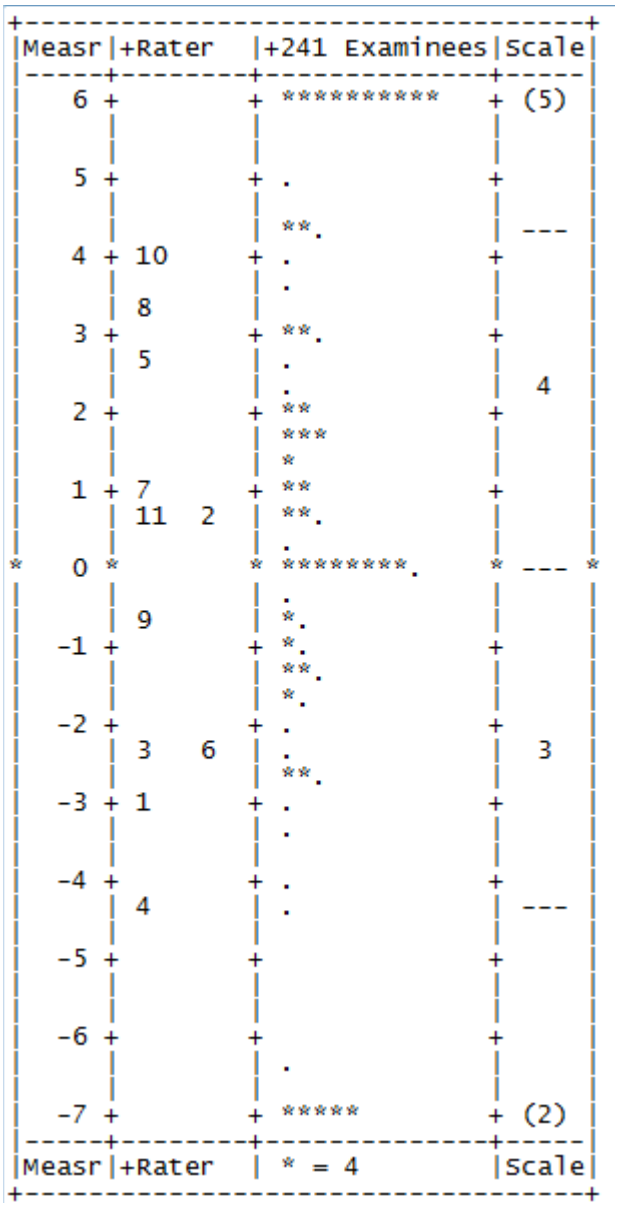

Figure 10. Criterion vowels level vertical scale.

Similarly to the criterion of word stress, the criterion of vowels was not a strong indicator to differentiating examinees' pronunciation ability, as shown by the examinees separation reliability of .42 or 17.6% of score variance. As with word stress, this statistic could mean that the rubric criterion descriptors need more refinement, the raters need more training or this could be an indication that the examinees have similar abilities in word stress.

Consonants. Unlike usage of the criterion of word stress, all the five categories under the criterion of consonants, had been used effectively by the raters. According to Table 10, the average measures advanced steadily without dropping, and the outfit mean-square statistics were less than 2.0. However, even though the threshold measures were more than 1.4 logits apart, Category 2 and 3 were more than 5 logits apart (see Figure 11), indicating a possibility that at least one of the two categories were overused. The differences between each threshold were evenly spaced, however, so this might be an artifact of wide logit range (e.g. -12 to 12). Lower categories being overused could indicate that consonants are hard in general for English learners.

Table 10

Criterion Consonants Rating Scale Category Statistics

Category	Absolute Frequency	Relative Frequency	Average Measure	Outfit	Threshold	SE
1	3	1%	-6.72	0.5	NA	NA
2	48	15%	-3.75	0.8	-8.07	0.75
3	110	35%	0.30	1.0	-2.34	0.26
4	136	43%	4.55	0.9	2.07	0.20
5	60	6%	8.28	1.1	8.34	0.34

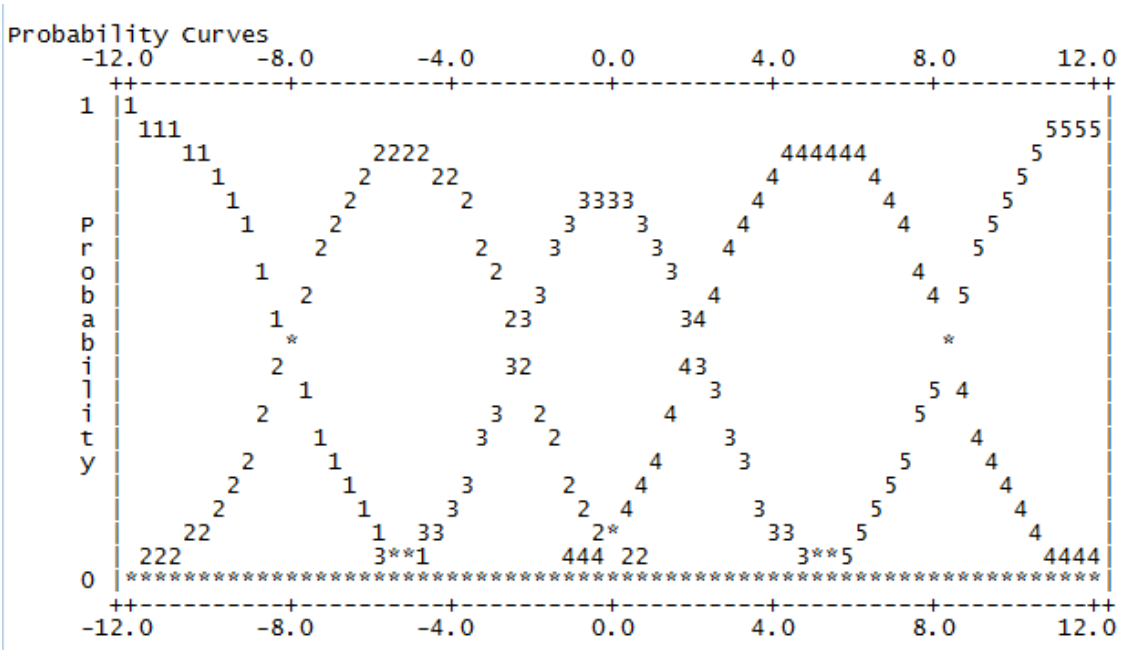

Figure 11. Probability curves of a given ability to produce consonants.

The analysis of the raters attained a separation reliability of .97, indicating that the raters had varying degrees of severity when performing the rating. From Figure 12, Rater 5 was the most generous and Rater 4 was the most severe. The fit statistics suggest that individual raters were consistent within themselves.

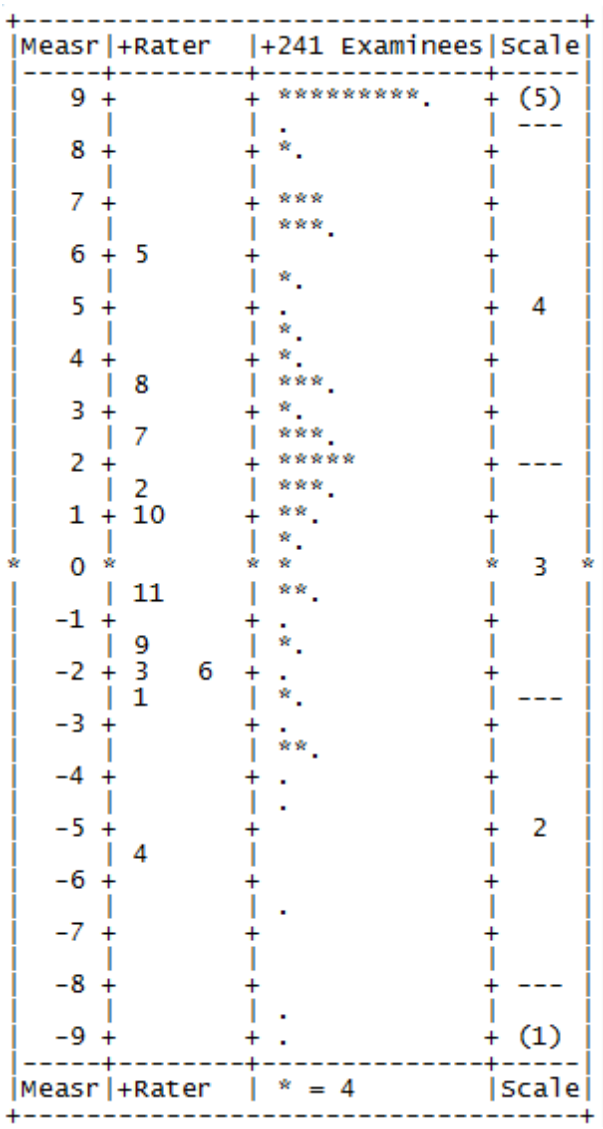

Figure 12. Criterion consonants level vertical scale.

The analysis resulted in an examinees separation reliability of .53, which means that 27.56% of the variance in different pronunciation ability could be accounted for by consonants alone. Taking into consideration that each criterion contributes its part to the overall pronunciation, the ratings of consonants seem to start differentiating examinee pronunciation ability.

Intonation. An analysis of the category statistics shows that the average measures and the threshold measures increased incrementally (see Table 11), which means higher ratings represent

higher ability in producing intonation. The outfit statistics did not exceed 2.0, which means that the average measures were within acceptable range. The differences between two adjacent threshold measures were larger than 5.0 logits; however, the differences between each threshold were evenly spaced. While this might be an indication that the categories were too large and could be split (Figure 13), it could be an artifact of wide logit range (e.g. -12 to 12).

Table 11

Criterion Intonation Rating Scale Category Statistics

Category	Absolute Frequency	Relative Frequency	Average Measure	Outfit	Threshold	SE
1	8	1%	-7.14	0.6	NA	NA
2	50	16%	-3.90	0.8	-8.00	0.69
3	121	38%	0.21	0.8	-2.52	0.25
4	128	40%	4.59	0.9	2.19	0.21
5	47	5%	7.70	1.1	8.33	0.35

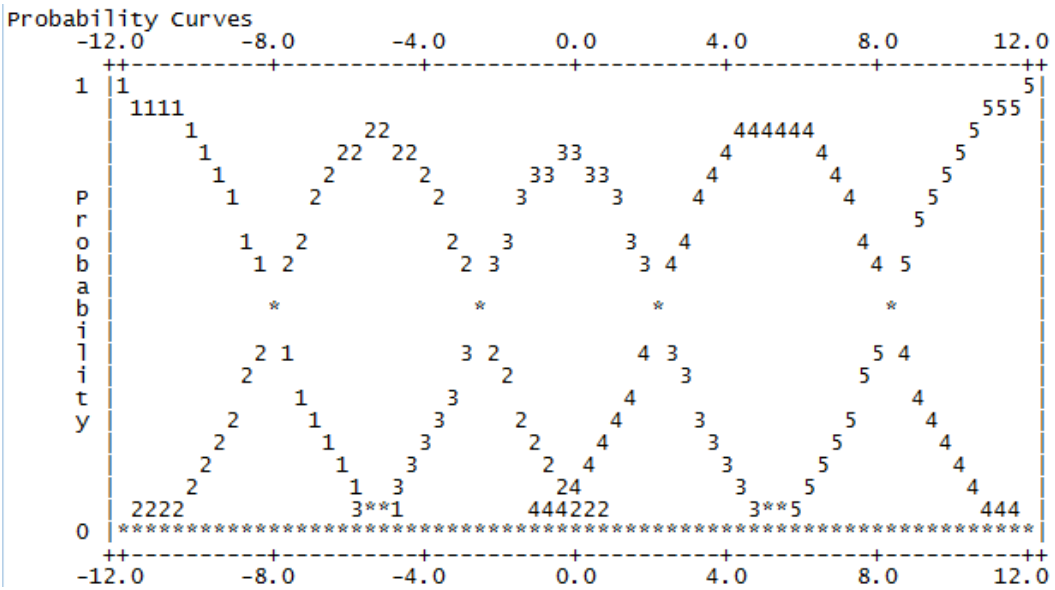

Figure 13. Probability curves of a given ability to produce intonation.

An analysis of the raters showed a separation reliability of was .97, which indicates that the raters had different severity in rating intonation. The raters, in general, used the rubric in an

internally consistent way. From Figure 14, most raters were severe when rating intonation with Rater 4 being the most severe.

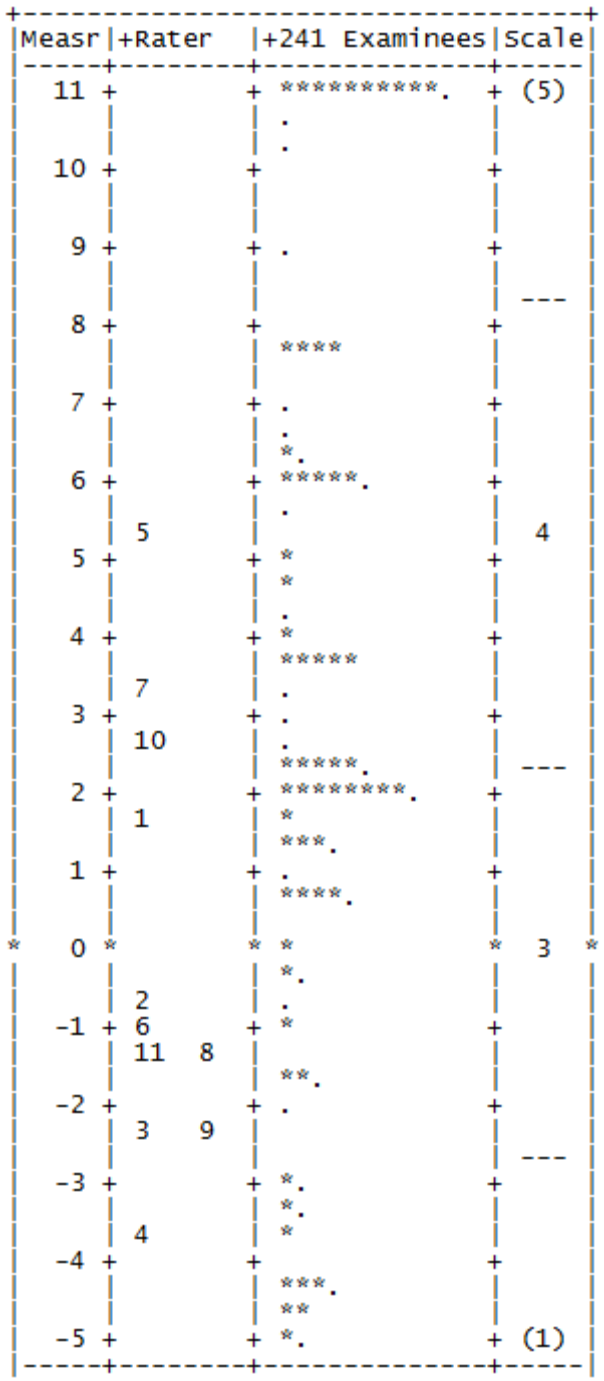

Figure 14. Criterion intonation level vertical scale.

An analysis of the examinees showed a separation reliability of .57, which is indicative that 32.49% of the variance in different pronunciation ability could be accounted for by intonation alone. As each criterion contributes to reliability of the overall pronunciation score, intonation appears to have greater effect than the segmentals.

Sentence Stress. The analysis showed that there were some undesirable category statistics for the rubric. From Table 12, the average measures did not increase incrementally. The average measure of Category 2 was lower than that of Category 1 and, the outfit statistic of Category 1 exceeded 2.0. These two pieces of information inform us that the difference between the expected measure and average measure of Category 1 seems problematic. However, upon further investigation, it is evident that there were fewer only 6 cases (less than 1%) in that category. The paucity of ratings in that category could have distorted both the measures and the fit statistics. In addition, the difference between Category 4 and 5 was 5.01 logits (see Figure 15), but the differences between each threshold were evenly spaced. While this might be an indication that the categories were too large and could be split, it could be an artifact of wide logit range (e.g. -9 to 9).

Table 12

Criterion Sentence Stress Rating Scale Category Statistics

Category	Absolute Frequency	Relative Frequency	Average Measure	Outfit	Threshold	SE
1	6	1%	-3.33	2.3	NA	NA
2	40	13%	-3.49*	1.0	-7.10	0.63
3	121	38%	0.22	0.7	-2.55	0.27
4	126	40%	4.37	0.8	2.32	0.20
5	63	8%	7.41	0.8	7.33	0.28

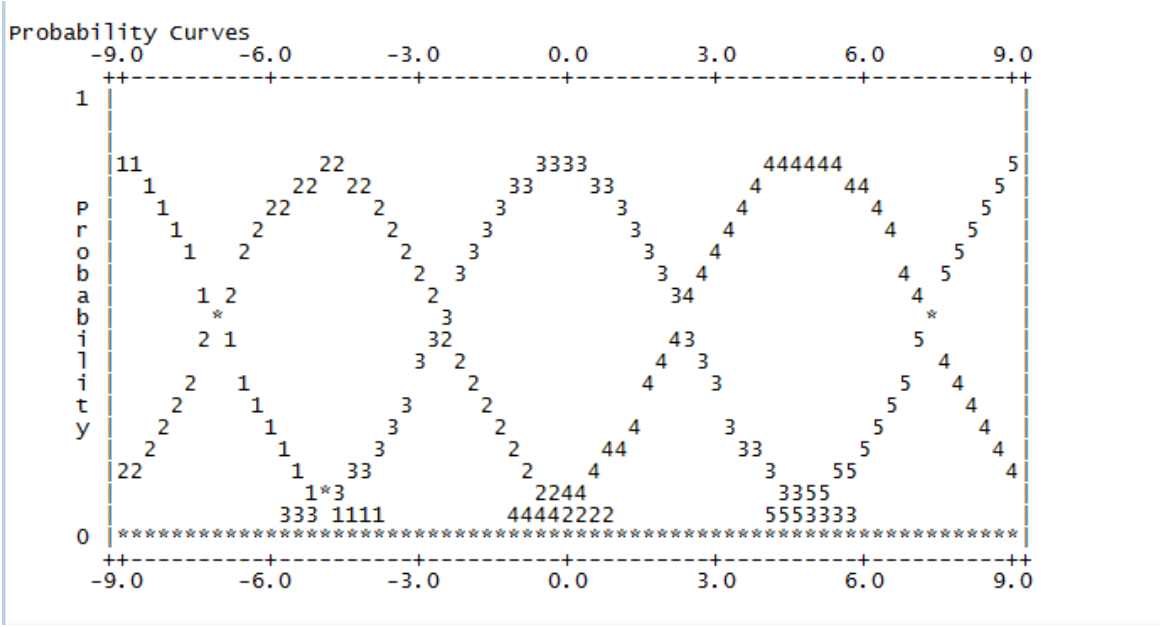

Figure 15. Probability curves of a given ability to produce sentence stress.

The raters were consistent in using the rating scale with exception of Rater 11 who had an outfit mean-square statistic of 1.90. The raters had different degrees of severity, indicated by a separation reliability of .95. Figure 16 visually presents the distribution of the examinees' ability and raters' severity. Rater 9 was the most severe and Rater 5 the most lenient.

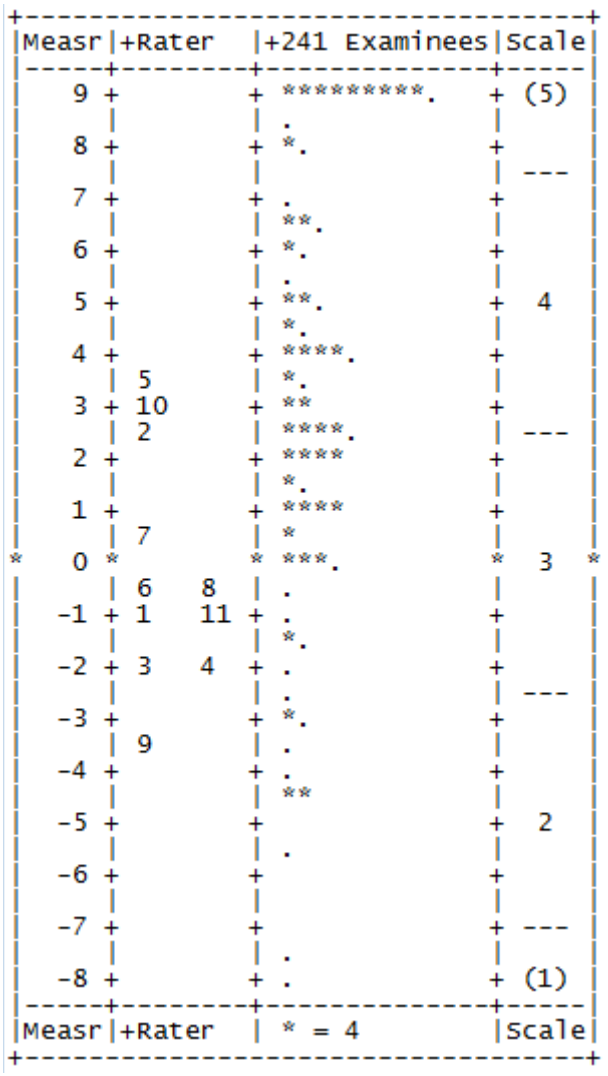

Figure 16. Criterion sentence stress level vertical scale.

The examinees' abilities in sentence stress ranged from Category 1 to Category 5 with the separation reliability between examinees being .63. This statistic indicates that sentence stress accounts for 39.6% of the score variance and can help differentiate examinees' abilities.

Rhythm. An analysis of the category statistics suggests that both the average measures and the threshold measures increased steadily without dropping (see Table 13), which is indicative that higher ratings represented higher ability of producing rhythm. The outfit statistics were less than 2.0, but the threshold measures of Category 4 and Category 5 were 5.21 logits apart, which was slightly more than 5.0 logits apart (see Figure 17) and while this might be an

indication that the categories were too large and could be split, it could be an artifact of wide logit range (e.g. -9 to 9).

Table 13

Criterion Rhythm Rating Scale Category Statistics

Category	Absolute Frequency	Relative Frequency	Average Measure	Outfit	Threshold	SE
1	13	1%	-4.13	1.4	NA	NA
2	52	17%	-3.93	0.6	-7.14	0.55
3	129	41%	0.18	0.8	-2.65	0.24
4	114	37%	4.15	1.0	2.29	0.21
5	49	4%	6.61	0.9	7.50	0.34

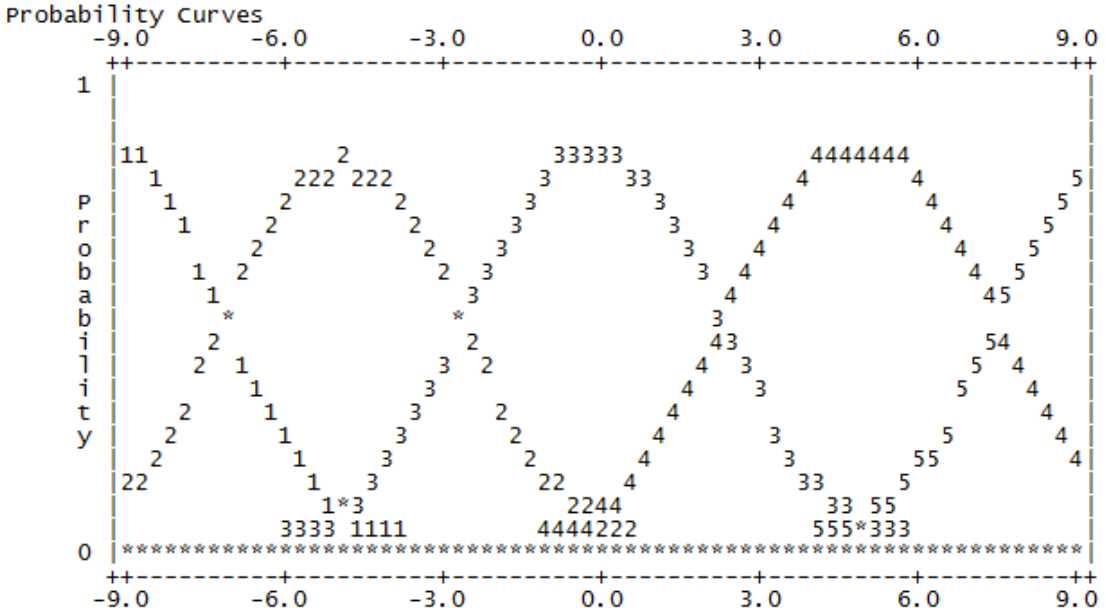

Figure 17. Probability curves of a given ability to produce rhythm.

All the raters except one used the rubric consistently internally, which was informed by the fit statistics being within a range of 0.5 to 1.5. Rater 11 demonstrated misfit with an outfit mean-square statistic of 1.62. The separation reliability was .93, meaning the raters had different degrees of severity in rating rhythm (see Figure 18).

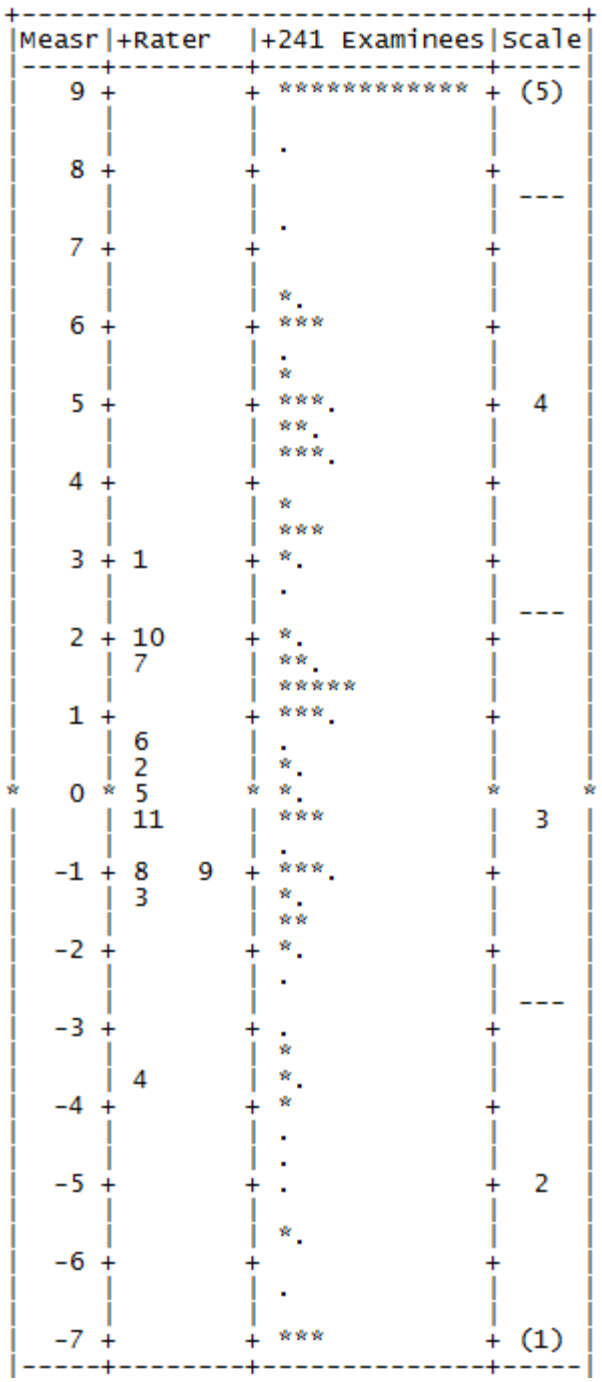

Figure 18. Criterion rhythm level vertical scale.

An analysis of the examinees showed a separation reliability of .68, showing that the 46.2 % of examinee score variance could be distinctly separated by rhythm. It indicates that the

examinees had different abilities in producing English rhythm and that the rubric and rating procedure were effective.

The analyses for the functionality and reliability of the pronunciation rubric at criterion level indicate that the rubric did not function as well at each criterion as it did as a whole. The reason for the rubric not ideally functioning at the micro level might be due to ability distribution of the examinees, insufficient training, and human errors which are unavoidable in every research setting and real life situation. However, global model fit should not be expected because the model was idealized. In research, the goal is not to achieve perfect model fit, but to find out what should be done.

Phase 2: Pronunciation features influence speaking proficiency. The question was examined of how well each of the pronunciation features in the pronunciation rubric predicts the overall speaking proficiency ratings. Table 14 reports the r-value, p-value, and N for each cell of the correlation of the correlation matrix. There were correlations between the speaking proficiency rating and the six explanatory variables, with the highest being sentence stress (r = .61). There were also inter-correlations among the six explanatory variables, with the highest being between sentence stress and rhythm (r = .72) and sentence stress and intonation (r = .70).

Previous research has indicated that suprasegmental features were more important than segmental features, so in the sequential regression analysis suprasegmentals were entered as independent variables before the segmentals. Also, because of the high correlation of sentence stress and two other variables, the explanatory variables were entered in order as: sentence stress, rhythm, intonation, word stress, consonants, and vowels. The first three models contained explanatory variables that best explain the response variable, speaking proficiency. The best model was Model 3, containing sentence stress, rhythm, and intonation, and nearly 42% of the

speaking proficiency rating (adjusted R^2 = .422) can be explained by these three variables. Table 15 shows R^2 for the model, adjusted R^2, and unstandardized regression coefficients (B) and their 95% CIs. This finding confirms other findings that suprasegmentals are more important (e.g. Kang, 2010) in determining overall speaking ratings. Even though word stress is a suprasegmental feature, adding it to the model does not improve the R square value. Therefore, better performance in sentence stress, rhythm, and intonation tends to lead to better proficiency ratings.

Table 14

Correlations Between Variables

		Proficiency	SentenceStress	Rhythm	Intonation	WordStress	Consonants	Vowels
Proficiency	r-value	1.000	.605	.596	.528	.430	.413	.436
	p-value	.	.000	.000	.000	.000	.000	.000
	N	221	221	221	221	221	221	221
SentenceStress	r-value		1.000	.720	.693	.631	.610	.609
	p-value		.	.000	.000	.000	.000	.000
	N		221	221	221	221	221	221
Rhythm	r-value			1.000	.615	.529	.547	.607
	p-value			.	.000	.000	.000	.000
	N			221	221	221	221	221
Intonation	r-value				1.000	.532	.551	.508
	p-value				.	.000	.000	.000
	N				221	221	221	221
WordStress	r-value					1.000	.571	.548
	p-value					.	.000	.000
	N					221	221	221
Consonants	r-value						1.000	.647
	p-value						.	.000
	N						221	221
Vowels	r-value							1.000
	p-value							.
	N							221

To further explore the influence of each of the six pronunciation features listed above, an error bar graph was plotted out (see Figure 19) with a 95% confidence interval. The levels in the graph were rounded values of the fair average of proficiency ratings. The first glance of this figure tells us that no examinees were rated to be placed in proficiency Level 1, which could be explained by the fact that the study was carried out at the end of the semester. The error bars at Level 7 were substantial because not very many people were placed at that level. As a result, the information gathered from Level 7 was less informative.

With regards to each pronunciation feature, there is a monotonical increase of ability from proficiency Level 2 to Level 6 in general, but some error bars overlap with that in the adjacent level. The overlap may indicate that this pronunciation feature does not serve to clearly distinguish the levels. It may also indicate that this pronunciation feature in learners change gradually as the proficiency increases. For example, there is an overlap of intonation between proficiency Level 3 and Level 4, which means a student who has a speaking proficiency at Level 3 and a student who has a speaking proficiency at Level 4 do not have distinct ability in intonation. Out of all the pronunciation features, only sentence stress has no major overlaps from level to level. Intonation had one overlap area between Level 3 and Level 4, rhythm had one overlap area between Level 2 and Level 3, and vowels had one overlap area between Level 2 and Level 3. Among the overlap areas, the top of the error bar of Level 2 of vowels exceeds that of Level 3, which means that a student with a lower proficiency could have higher ability in vowels. Therefore, after sentence stress, intonation and rhythm best predict speaking proficiency, verifying the statistical results reached above.

Table 15

Summary of Regression Models to Predict Speaking Proficiency

Model	R^2	Adjusted R^2	Sentence Stress B	Rhythm B	Intonation B	Word Stress B	Consonants B	Vowels B
1	.366	.363	.812 (.669, .954)					
2	.420	.415	.489 (.293, .686)	.421 (.236, .605)				
3	.430	.422	.390 (.170, .610)	.378 (.189, .566)	.198 (-.003, .399)			
4	.430	.420	.377 (.144, .611)	.375 (.185, .564)	.193 (-.010, .397)	.030 (-.161, .222)		
5	.430	.417	.383 (.145, .621)	.378 (.186, .570)	.198 (-.009, .404)	.037 (-.161, .236)	-.027 (.232, .178)	
6	.430	.414	.381 (.141, .621)	.374 (.176, .572)	.198 (-.009, .404)	.035 (-.166, .237)	-.033 (-.254, .187)	.018 (-.213, .249)

a. Predictors: (Constant), SentenceStress
b. Predictors: (Constant), SentenceStress, Rhythm
c. Predictors: (Constant), SentenceStress, Rhythm, Intonation
d. Predictors: (Constant), SentenceStress, Rhythm, Intonation, WordStress
e. Predictors: (Constant), SentenceStress, Rhythm, Intonation, WordStress, Consonants
f. Predictors: (Constant), SentenceStress, Rhythm, Intonation, WordStress, Consonants, Vowels
g. Dependent Variable: Proficiency

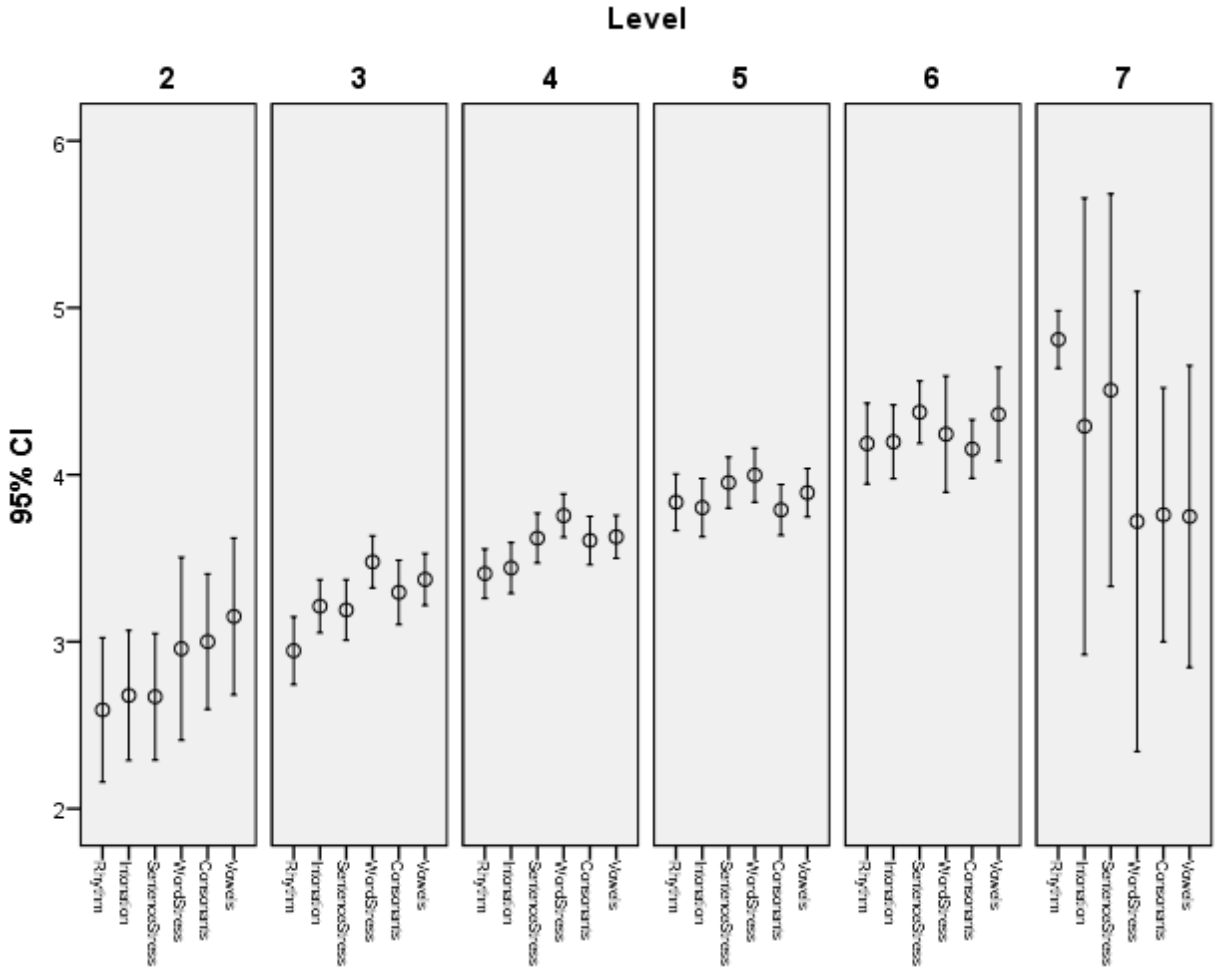

Figure 19. Error bar graph of each pronunciation feature at each proficiency level.

Conclusion

The purpose of this thesis was to explore the role of pronunciation and its particular features (i.e. vowels, consonants, intonation, word stress, rhythm, and sentence stress) in determining overall speaking test ratings. The data were collected at BYU's ELC where raters rated each pronunciation feature as well as overall speaking proficiency using MFRM. The ratings were calculated into fair average scores, and these scores were used in regression analyses to answer the research questions.

The results from the analyses show that pronunciation, especially sentence stress, intonation, and rhythm, could explain nearly 42% of the variance of speaking proficiency ratings. Among the pronunciation features listed above, sentence stress was the factor that contributed the most to explaining proficiency ratings.

This section will elaborate on the results described in the preceding section. The data resulting from the ratings on vowels, consonants, intonation, word stress, rhythm, and sentence stress are contextualized in terms of classroom practice and curriculum design. Additionally, some limitations of the current study are addressed, and possibilities for future research are proposed.

Pedagogical Implications

From the current study, it is safe to conclude that pronunciation is crucial in determining a learner's speaking proficiency level. It is important to call educators' attention to teaching pronunciation and to giving pronunciation related feedback and instruction that could assist learners in achieving a higher level of proficiency.

The findings confirmed that suprasegmentals are more important in achieving a high proficiency level in speaking than segmentals. Even though the research has enlarged the

importance of suprasegmentals, the practice in classrooms hardly is the realization of the belief. Studies done in Canada showed that the beliefs of the teachers were that they emphasized suprasegmentals, but relied on segmental-based materials (Breitkreutz, Derwing, & Rossiter, 2001). A decade later, the instruction in pronunciation has not changed sufficiently (Foote, Holtby, & Derwing, 2011). The current study suggests approaching pronunciation teaching through suprasegmental instruction as J. B. Gilbert (2008) has proposed in her Prosody Pyramid.

Among the suprasegmental features, it is suggested that substantial attention be given to sentence stress in pronunciation teaching and that it be taught at all proficiency levels. The current study shows that sentence stress plays a larger role in determining speaking proficiency levels. This finding coincides with the studies by Bansal (1969) who analyzed Indian English and by Hahn (2004) who examined content recall from three formats of an ITA's lecture. These studies serve as evidence of the importance of suprasegmentals in communication. The findings of the current study also show that sentence stress is distinct at each speaking proficiency level. Therefore, as the learners' overall proficiency improves, they should continue receiving instruction in sentence stress.

In summary, pronunciation is essential in achieving a high level of speaking proficiency, and sentence stress is the most important factor among the pronunciation features included in this study, so English teachers need to incorporate sentence stress instruction.

Limitations

As with other studies, this study does have some limitations. Directing attention to these limitations leads to the appropriate use of the results. The limitations include speaking sample elicitation, rating procedure, and the analysis.

Speaking sample elicitation. The speaking samples used in the current study were obtained from a proficiency test. The students took their test so that their study level of the following semester could be determined. Accordingly, in order to be rated on their highest performance, the examinees might have avoided producing certain sounds on the test. For example, an examinee with difficulties in saying /i/ and /i:/ might have avoided the /i/ sound and instead chose to use the word "paper" instead of the word "sheets." Such compensation strategies may disguise students' pronunciation errors, which may have caused students to receive higher pronunciation scores and may not have reflected their true pronunciation ability. However, the effect of compensation strategies is not clear in the current study or in proficiency ratings.

Data collection. In the process of data collection, there may have been interactions among rating criteria and interactions between rater and examinee. The proficiency and each criterion of pronunciation were rated at the same time. There was an unavoidable interaction between the proficiency rating and the pronunciation rating. Some raters may have been unwilling to give low pronunciation ratings to a highly proficient sample. Also, there could have been interactions among the six pronunciation criteria. Rating all six features at once could have been overwhelming for raters to switch among the criteria. Some raters may also have demonstrated rater-examinee interaction error, which might cause them to give low pronunciation ratings to highly proficient speaking samples if they were not familiar with the accent.

Also, in the current study, the raters chose the speaking samples that, in their own opinions, represented the examinee's pronunciation. Even though the reason behind this practice was to avoid the constraint of the prompt on the examinee's performance, raters choosing different responses might end up decreasing the validity of the study.

The rubric use on criterion level. Even though the rubric functioned ideally holistically, it did not effectively perform its function perfectly at the criterion level. All six criteria had some statistics that were not within the acceptable range. Take the outlier rater, Rater 7, for example. The overall analysis shows that the behavior of this rater was completely within an acceptable range, but this rater consistently had more variation than expected when rating all criteria except one. The reason could be that the raters were using the rubric for the first time with little training. One rater commented that if there had been benchmark samples they could listen to, they would have been more confident in using the rubric.

Implications for Future Research

The results of this study lead to several suggestions for future research. Improvements could be made on this study to help generalize the results, and further research could be conducted to explore similar issues in other English speaking proficiency tests.

In order to alleviate the current study, several suggestions are given. One suggestion to improve the current study would be to further minimize rater-examinee interactions. One way to do that would be to recruit more raters who could not speak Spanish and assign speaking samples produced by Spanish speakers to them. In the current study, Spanish speaking learners and raters were the majority. There were unavoidable rater-examinee interactions among the Spanish speaking English learners and the raters. This interaction might have influenced the results of this study, but it was not analyzed. Further research could take measures to avoid the interaction effect through recruiting more raters who do not speak Spanish.

A further suggestion to improve the current study would be in the aspects of the speaking sample elicitation method and the data collection method. The researcher could determine which responses should be listened to for the pronunciation rating. In the current study, the raters were

given the freedom of choosing the responses to listen to, after they performed the proficiency rating. There may have been invalidity brought by ratings from different responses. Also, if time and the program permit, the raters could rate each pronunciation feature and the proficiency separately. The validity may be ameliorated because the raters would be focused on one trait at a time.

To contribute to pronunciation and proficiency related research in the future, some suggestions are possible. One suggestion relates to the rubric. As reported in the results section, the category statistics and fit statistics showed undesirable use of the rubric. The data could be recoded to generate more valid results. An alternative would be to further refine the rubric. The pronunciation rubric used in this study was not perfect. The descriptors, though detailed, were not research based. Based on the statistics from this study, the scale categories of vowels could be collapsed, and the scale categories of intonation and rhythm could be expanded. The statistics also indicated that the descriptor of Category 1 in word stress might not be an accurate description of the lowest performance because speaking samples were rarely placed in that category. The descriptors of Category 1 and Category 2 in sentence stress might not have distinguishable enough difficulties because Category 2 had a slightly lower average measure than that of Category 1. In order to find out how to improve the rubric, qualitative data could be gathered from raters through interviews or thinking aloud. Information could be collected, such as the way the raters interpret the rubric, the strengths and weaknesses of the rubric, and ease of using the rubric.

In addition to refining the rubric, more rater training could be provided so that the raters could feel more confident in using the rubric to rate pronunciation. The effectiveness of rating training in the current study was not measured. Even though the calibration was designed to

bring raters to a similar understanding of the rubric and to improve the use of the rubric, raters still used the rubric in a number of ways indicated by the rater separation strata and separation reliability. The training could include providing benchmark samples for the raters and discussing thoughts about those samples based on the rubric.

Summary

To summarize the findings of the study, pronunciation could explain 42% of variance of speaking proficiency ratings, and sentence stress was the most important factor in determining the overall speaking proficiency rating. The results gained in this study may not only be beneficial for English teachers and English learners, but they could also provide insight for curriculum development and policy making regarding pronunciation teaching and learning.

References

American Council on the Teaching of Foreign Languages. (2012). ACTFL Proficiency Guidelines 2012. Retrieved from http://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

Acton, B. (2015). Acton Haptic-English pronunciation system [Online post]. Retrieved from http://www.actonhaptic.com

Adams, C. (1979). *English speech rhythm and speech patterns.* The Hague: Mouton.

Bansal, R. (1969). *The intelligibility of Indian English.* Hyderabad, India: Central Institute of English.

Bowen, J. D. (1972). Contextualizing pronunciation practice in the ESOL classroom. *TESOL Quarterly, 6*, 83-94.

Bowen, J. D. (1975). *Patterns of English pronunciation.* Rowley, MA: Newbury House.

Brazil, D., (1997). *The communicative value of intonation in English.* Cambridge: Cambridge University Press.

Brazil, D., Coulthard, M., & Johns, C. (1980). *Discourse intonation and language teaching*. London: Longman Group.

Breitkreutz, J. A., Derwing, T. M., & Rossiter, M. J. (2001). Pronunciation teaching practices in Canada. *TESL Canada Journal, 19*(1), 51-61.

British Council, IDP: IELTS Australia, & UCLES. (n.d.) *IELTS Speaking Band Descriptors (Public Version)*. Retrieved from

https://www.teachers.cambridgeesol.org/ts/digitalAssets/114292_IELTS_Speaking_Band_Descriptors.pdf

Brown, A. (1988). Functional load and the teaching of pronunciation. *TESOL Quarterly, 22*, 593-606.

Canagarajah, S. (2014). In search of a new paradigm for teaching English as an international language. *TESOL Journal, 5*. 767-785. doi:10.1002/tesj.166

Canale, M.; Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2010). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*, 201-219. doi: 10.1177/0265532210393704

Celce-Murcia, M. (1987). Teaching pronunciation as communication. In J. Morley (Ed.), *Current perspectives on pronunciation: Practices anchored in theory* (pp.5-12). Washington, DC: TESOL.

Celce-Murcia, M. (2013). Teaching English in the context of world Englishes. In M. Celce-Murcia, D. M. Brinton & M. A. Snow (Eds.), *Teaching English as a Second or Foreign Language* (4ed, pp. 2-14). Boston, MA: National Geographic Learning/Cengage Learning.

Celce-Murcia, M., Brinton, D., & Goodwin, J. (2010) *Teaching pronunciation: A reference for teachers of English to speakers of other languages.* Great Britain: Oxford University Press.

Celce-Murcia, M. & Olshtain, E. (2013). Teaching language through discourse. In M. Celce-Murcia, D. M. Brinton & M. A. Snow (Eds.), *Teaching English as a Second or Foreign*

Language (4ed, pp. 424-437). Boston, MA: National Geographic Learning/Cengage Learning.

Council of Europe. (n.d.). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Retrieved from http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf

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

Dalton, P. & Hardcastle, W. J. (1977). *Disorders of fluency and their effects on communication.* London: Edward Arnold Ltd.

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

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

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

Derwing, T. M., Munro, M. J., & Wiebe, G. (1997). Pronunciation instruction for "fossilized" learners: Can it help? *Applied Language Learning, 8*, 217-235.

Douglas, D. (1994). Quantity and quality in speaking test performance. *Language Testing, 11*, 125-144. doi: 10.1177/036553229401100203

Educational Testing Service (2004). *TOEFL iBT/Next Generation TOEFL Test*. Retrieved from http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf

Educational Testing Service (2006). Rubric for rubrics. Retrieved from https://www.mbaea.org/documents/filelibrary/pdf/csin/RubricforRubrics_77EAE6205D215.pdf

Evans, N. W., Hartshorn, K. J., Cox, T. L., & de Jel, T. M. (2014). Measuring written linguistic accuracy with weighted clause ratios: A question of validity. *Journal of Second Language Writing, 24*, 33-50.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments.* Frankfurt am Main: Peter Lang.

Fayer, J. M. & Krasinski, E. (1987). Native and non-native judgments of intelligibility and irritation. *Language Learning, 37*, 313-326.

Foote, J. A., Holtby, A. K., Derwing, T. M. (2011) Survey of the teaching of pronunciation in adult ESL programs in Canada. *TESL Canada Journal, 29*(1), 1-22.

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

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 233-277). Timonium, MD: York Press.

Gallego, J. C. (1990). The intelligibility of three nonnative English-speaking teaching assistants: An analysis of student reported communication breakdown. *Issues in Applied Linguistics. 1,* 219-237.

Gilbert, A. C. (2014). *The perceptual chunking of speech: On the nature of temporal grouping and its effect on immediate memory* (Doctoral dissertation). Retrieved from ProQuest. (Order No. AAINR79277)

Gilbert, J. B. (2008). *Teaching pronunciation: Using the prosody pyramid.* New York, NY: Cambridge University Press.

Gilbert, J. B. (2012). *Clear speech*. New York, NY: Cambridge University Press

Goodwin, J. (2013). Teaching pronunciation. In M. Celce-Murcia, D. M. Brinton & M. A. Snow (Eds.), *Teaching English as a Second or Foreign Language* (4ed, pp. 136-152). Boston, MA: National Geographic Learning/Cengage Learning.

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*, 201-223.

Henrichsen, L.E. (n.d.a). ESL 302 advanced English pronunciation oral presentation evaluation rubric. Retrieved from personal communication.

Henrichsen, L.E. (n.d.b). Biography [biography]. Retrieved from http://linguistics.byu.edu/directory/leh5/

Higgs, T. V. & Clifford, R. (1982). The push toward communication. In T. V. Higgs (Ed.), *Curriculum, competence and the foreign language teacher* (pp. 51-79). Lincolnwood, IL: National Textbook.

Isaacs, T. & Trofimovich, P. (2011). Phonological memory, attention control, and musical ability: Effects of individual differences on rater judgments of second language speech. *Applied Psycholinguistics, 32*, 113-140.

Isaacs, T. & Trofimovich, P. (2012). Deconstructuring comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34,* 475-505. doi: 10.1017/S0272263112000150

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an International Language. *Applied Linguistics, 23*, 83-103.

Jenkins, J. (2006). Current perspectives on teaching World Englishes and English as a Lingua Franca. *TESOL Quarterly, 40*, 157-181.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System, 38*, 301-315. doi: 10.1016/j.system.2010.01.005

Kang, O. (2013). Relative impact of pronunciation features on ratings of non-native speakers' oral proficiency. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference* (Aug. 2012. pp. 10-15). Ames, IA: Iowa State University.

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

Lee, Y. J. (2012). Software to facilitate language assessment: Focus on Quest, Facets, and Turnitin. In C. Coombe, P. Davidson, B. O'Sullivan, & S., Stoynoff (Eds.), *The* Cambridge Guide to second language assessment (pp. 280-288). New York, NY: Cambridge University Press.

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*, 369-377.

Levis, J. M. (2006). Pronunciation and the assessment of spoken language. In R. Hughes (Ed.), *Spoken English, TESOL and applied linguistics: Challenges for theory and practice* (pp. 245 – 270). New York: Palgrave Macmillan.

McNerney, M., & Mendelsohn, D. (1992). Suprasegmentals in the pronunciation class: Setting priorities. In P. Avery & S. Ehrlich (Eds.), *Teaching American English Pronunciation* (pp. 185–196). Oxford: Oxford University Press.

Morley, J. (1991). The pronunciation component in teaching English to speakers of other languages, *TESOL Quarterly, 25*, 481-520.

Mochizuki-Sudo, M., & Kiritani, S. (1991). Production and perception of stress-related durational patterns in Japanese learners of English. *Journal of Phonetics, 19*, 231–248.

Munro, M. J., & Derwing, T.M. (1995). Processing time, accent and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*, 289-306.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *49*, 285-310.

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

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

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Nicolosi, L., Harryman, E., & Keresheck, J. (1989). *Terminology of communication disorders.* Baltimore, MD: William & Wilkins.

Pennington, M. C., & Ellis, N. C. (2000). Cantonese speakers' memory for English sentences with prosodic cues. *The Modern Language Journal, 84*, 372-389.

Pennington, M. C., & Richards, J. C. (1986). Pronunciation revisited. *TESOL Quarterly, 20*, 207–225.

Pickering, L. (2001). The role of tonal choice in improving ITA communication in the classroom. *TESOL Quarterly, 35*, 233-255.

Riney, T. J., Takagi, N. & Inutsuka, K. (2005). Phonetic parameters and perceptual judgments of accent in English by American and Japanese listeners. *TESOL Quarterly, 39*, 441-466.

Rose, M. (n.d.). Make Room for Rubrics. Retrieved from http://emp.byui.edu/firestonel/bio405/readings/assessmemt/make%20room%20for%20rubrics.pdf

Sifakis, N. C. & Sougari, A., (2005). Pronunciation issues and EIL pedagogy in the periphery: A study of Greek State School teachers' beliefs. *TESOL Quarterly, 39*, 467-488.

Szpyra-Kozlowska, J., Frankiewicz, J., Nowacka, M., & Stadnicka, L. (2005). Assessing assessment methods - on the reliability of pronunciation tests in EFL. Retrieved from file:///Users/elcteacher/Desktop/ptlcp37.pdf

Taylor, K. & Thompson, S. (2013). Color vowel chart [Online Post]. Retrieved from http://colorvowelchart.org/

Watanabe, K. (1988). Sentence stress perception by Japanese students. *Journal of Phonetics, 16*, 181–186.

Wenk, B. (1985). Speech rhythms in second language acquisition. *Language and Speech, 28*, 157–175.

Wennerstrom, A. (1994). Intonational meaning in English discourse: A study of non-native speakers. *Applied Linguistics, 15*, 399–420.

Wennerstrom, A. (1997). Discourse intonation and second language acquisition: Three genre based studies (Unpublished doctoral dissertation). University of Washington, Seattle.

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

Zielinski, B. W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System, 36*, 69-84. doi: 10.1016/j.system.2007.11.004

Appendix A

Speaking Proficiency Test Prompts

1. [Warm-up] (15 prep/30 speak) Describe the weather and scenery as you came to take this test today. You have 15 seconds to prepare and 30 seconds to speak.

2. [Novice] (15 prep/30 speak) Describe what you are wearing today. List the clothes and identify their color, material, and other characteristics. Also include your reason for choosing to wear them. You have 15 seconds to prepare and 30 seconds to speak.

3. [Intermediate] (15 prep/45 speak) Do your best to describe where you will be and what you will be doing one year in the future. How will your life be different? How will it be the same? What events will happen between now and one year from now? You have 15 seconds to prepare and 45 seconds to speak.

4. [Intermediate - Questions] (15 prep/30 speak) You and your classmates want to plan a party for one of your teachers who is moving after the semester. What are several questions you should ask your teacher in order to plan a party that she would like? You have 15 seconds to prepare and 30 seconds to speak.

5. [Advanced – Description] (30 prep/60 speak) Describe a holiday in your country that the U.S. does not celebrate. What is the reason for the holiday? How do people celebrate? What are things that a person would see, do or eat if they visited your country during that holiday? You have 30 seconds to prepare and 60 seconds to speak.

6. [Superior - Support an opinion] (30 prep/90 speak) Two friends are having a debate. One friend believes that playing video games is a waste of time and parents should prohibit their use. The other friend believes that children can acquire valuable skills from video games and

parents should facilitate their use. Choose one side of this argument to support and explain your reasons for having your opinion. You have 30 seconds to prepare and 90 seconds to speak.

7. [Advanced – Resolution of problem] (30 prep/60 speak) You are working with a group of classmates to complete a presentation. Your responsibility was to create a media presentation with information and pictures that other group members researched. On the day of the presentation, you lose the USB drive containing the presentation and all of the information the group had collected. Explain to your group members what happened and describe a series of actions that the group should do to reach the best result. You have 30 seconds to prepare and 60 seconds to speak.

8. [Superior – Hypothesis] (30 prep/90 speak) In many countries, people are moving from rural areas into urban areas. Discuss the short term and long term consequences of this type of population movement.

9. [Superior – Abstract Discussion] (30 prep/90 speak) Edwin Land. an American inventor, said.

"An essential aspect of creativity is not being afraid to fail."

Discuss the principle behind this expression. In what way is it true or accurate? Who should learn from it and how should their actions change? You have 30 seconds to prepare and 90 seconds to speak.

10. [Advanced – Narration] (30 prep/60 speak) Retell a story from your life when you or someone you know won a prize or award. Include a detailed description of the events before, during and after this experience. How or why was this experience memorable to you? You have 30 seconds to prepare and 60 seconds to speak.

11. [Intermediate – Create] (15 prep/45 speak) A friend from your hometown asks about what you do on the weekend now that you live in the US. Describe your routine on a typical Saturday from the morning to the evening. What do you do? Where do you go? Who are you with? How is it different than weekends in your hometown? You have 15 seconds to prepare and 45 seconds to speak.

12. [Cool-down] (15 prep/30 speak) What are your plans for the rest of the day? What will you do to relax and enjoy the time following your test? You have 15 seconds to prepare and 30 seconds to speak.

Appendix B

Level	Text Type	Content	Accuracy
	• **Fluency** • **Development** • **Organization**	• **Functional Ability with the Language (Abstract vs. Concrete or Self-centric Language)** • **Vocabulary**	• **Grammar & Verb Tense** • **Communication Strategies** • **Native-like Comprehensibility**
7-ready for university courses	Exemplified speaking on a paragraph level rather than isolated phrases or strings of sentences. Highly organized argument (transitions, conclusion, etc.). Speaker explains the outline of topic and follows it through.	• Discusses some topics abstractly (areas of interest or specific field of study); • Better with a variety of concrete topics; • Appropriate use of a variety in academic and non-academic vocabulary;	• Grammar errors are extremely rare, if they occur at all; wide range of structures in all time frames; • Able to compensate for deficiencies by use of communicative strategies—paraphrasing, circumlocution, illustration—such that deficiencies are unnoticeable; • Readily understood by native speakers unaccustomed to non-native speakers;
6-ready for Academic C	Fairly organized paragraph-like speech with appropriate discourse markers (transitions, conclusion, etc.) Will not be as organized as level 7, but meaning is clear.	• Can speak comfortably with concrete topics, and discuss a few topics abstractly; • Academic vocabulary often used appropriately in speech;	• Grammar errors are infrequent and do not affect comprehension; no apparent sign of grammatical avoidance; • Able to speak in all major time frames, but lacks complete control of aspect; • Often able to successfully use compensation strategies to convey meaning; • Easy to understand by native speakers unaccustomed to non-native speakers
5-ready for Academic B	Simple paragraph length discourse with sustained, though possibly formulaic, discourse markers that help maintain some organization.	• Able to comfortably handle all uncomplicated tasks relating to routine or daily events and personal interests and experiences; • Some hesitation may occur when dealing with more complicated tasks; • Uses a moderate amount of academic vocabulary;	• Uses a variety of time frames and structures; however, speaker may avoid more complex structures; • Error patterns may be evident, but errors do not distort meaning; • Exhibits break-down with more advanced tasks—i.e. failure to use circumlocution, significant hesitation, etc. • Understood by native speakers unaccustomed to dealing with non-natives, but 1st language is evident;
4-ready for Academic A	Uses moderate-length sentences with simple transitions to connect ideas. Sentences may be strung together, but may not work together as cohesive paragraphs.	• Able to handle a variety of uncomplicated tasks with concrete meaning; • Expresses meaning by creating and/or combining concrete and predictable elements of the language; • Uses sparse academic vocabulary appropriately;	• Strong command of basic structures; error patterns with complex grammar; • Frequent use of compensation strategies with varied success; • Generally understood by sympathetic speakers accustomed to speaking with non-natives;

Level	Text Type	Content	Accuracy
	• **Fluency** • **Development** • **Organization**	• **Functional Ability with the Language (Abstract vs. Concrete or Self-centric Language)** • **Vocabulary**	• **Grammar & Verb Tense** • **Communication Strategies** • **Native-like Comprehensibility**
3-ready for Foundations C	Able to express personal meaning by using simple, but complete, sentences they know or hear from native speakers.	Able to successfully handle a limited number of uncomplicated tasks; Concrete exchanges and predictable topics necessary for everyday life without unexpected complications; Highly varied general vocabulary;	Errors are not uncommon and sometimes obscure meaning; Limited range of sentence structure; Characterized by ineffective reformulations and self-corrections; Generally understood by speakers used to dealing with non-natives, but requires more effort;
2—ready for Foundations B	Short and sometimes incomplete sentences.	Restricted to a few of the predictable topics necessary for survival (basic personal information, basic objects, preferences, and immediate needs) Relies heavily on learned phrases or recombination of phrases and what they hear from interlocutor; Limited general vocabulary	Attempt to create simple sentences, but errors predominate and distort meaning; Avoids using complex structures. Speaker's 1st language strongly influences syntax; Generally understood by sympathetic speakers used to non-natives with repetition and rephrasing;
1-ready for Foundations A	Isolated words and memorized phrases.	Relies almost solely on formulaic/memorized language; Two or three word answers in responding to questions; Very limited context for vocabulary;	Communicate minimally and with difficulty; Frequent pausing, recycling their own or interlocutor's words; Resort to repetition, words from their native language, or silence if task is too difficult; Understood with great difficulty even by those used to dealing with non-natives;
0-ready for Foundations prep	Isolated words.	No real functional ability; Given enough time and familiar cues, may be able to exchange greetings, give their identity and name a number of familiar objects from their immediate environment;	Cannot participate in true conversational exchange; Length of speaking sample may be insufficient to assess accuracy; Nearly incomprehensible even by those used to dealing with non-natives.

Appendix C

Pronunciation Rating Handout

In this pronunciation rubric, each of a number of ordered categories represents a successively higher level of performance. There are five categories. Category 5 is NOT native level.

- Definitions of terms used in the pronunciation rubric:

Intonation: "pattern of pitch and stress in the flow of speech" (Nicolosi, Harryman, & Kresheck, 1989, p.134). Speakers convey their emotions through intonation (Wennerstrom, 1997). English speakers use rising tones to avoid the appearance of overt disagreement, to review, and to indicate the assumption that the listeners already knew.

Word stress: "amount of force or strength of movement in the production of one syllable as compared with another; usually results in the syllable sounding longer and louder than other syllables in the same word" (Nicolosi, Harryman, and Kresheck, 1989, p.250).

Rhythm: "a temporal sequencing of similar events (Dalton and Hardcastle, 1977, p.41)." They explained that the "similar events" could be recurring patterns of more salient syllables than adjacent ones.

Sentence stress: Sentence stress is also known as primary stress, using to draw attention to new or contrastive information.

Vowel dense contexts: several minimal pairs appear in the same sentence or very close to each other, e.g. Matt has a bad bed.

She sighed, "Apples are gone."

- Consonant and vowel common errors in this rubric were added based on their functional load. According to how important they are in English communication and how easily they hinder communication, minimal pairs are listed in the rubric.

Rater 1

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	2	3	3	3	3	2	2.67
0431	2	2	3	3	3	4	2.83
0684	4	4	4	5	5	5	4.50
1568	3	3	3	2	3	4	3.00
2094	3	3	4	4	4	5	3.83
3824	4	4	2	3	3	4	3.33
7375	4	4	4	3	4	4	3.83
7633	3	2	2	3	4	4	3.00
8473	3	3	3	3	2	4	3.00
9031	2	2	3	3	3	3	2.67
9616	4	4	4	4	5	5	4.33
average	3.09	3.09	3.18	3.27	3.55	4.00	

Average

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3.00	2.89	3.22	2.89	2.56	2.44	2.83
0431	2.56	2.67	3.00	2.89	2.44	2.33	2.65
0684	3.89	4.00	4.22	4.33	4.33	4.22	4.17
1568	3.22	3.11	3.33	2.78	3.00	2.89	3.06
2094	3.67	3.78	4.22	4.00	3.89	4.00	3.93
3824	4.00	3.56	4.00	3.89	3.56	3.44	3.74
7375	4.00	3.78	4.00	4.00	3.89	3.78	3.91
7633	3.00	2.89	3.33	3.33	3.33	3.00	3.15
8473	3.22	3.00	3.33	3.00	3.44	3.00	3.17
9031	2.78	2.67	3.56	3.11	3.00	2.89	3.00
9616	4.44	4.56	4.67	4.44	4.33	4.56	4.50
average	3.43	3.35	3.72	3.52	3.43	3.32	

Rater 3

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3	3	3	3	3	2	2.83
0431	3	3	3	3	2	2	2.67
0684	4	3	4	4	4	4	3.83
1568	2	3	3	3	3	3	2.83
2094	4	3	4	4	4	4	3.83
3824	3	3	4	4	4	4	3.67
7375	3	4	4	4	4	4	3.83
7633	3	2	3	4	4	4	3.33
8473	3	3	4	3	4	3	3.33
9031	3	3	4	3	3	3	3.17
9616	4	4	5	4	4	4	4.17
average	3.18	3.09	3.73	3.55	3.55	3.36	

Average

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3.00	2.89	3.22	2.89	2.56	2.44	2.83
0431	2.56	2.67	3.00	2.89	2.44	2.33	2.65
0684	3.89	4.00	4.22	4.33	4.33	4.22	4.17
1568	3.22	3.11	3.33	2.78	3.00	2.89	3.06
2094	3.67	3.78	4.22	4.00	3.89	4.00	3.93
3824	4.00	3.56	4.00	3.89	3.56	3.44	3.74
7375	4.00	3.78	4.00	4.00	3.89	3.78	3.91
7633	3.00	2.89	3.33	3.33	3.33	3.00	3.15
8473	3.22	3.00	3.33	3.00	3.44	3.00	3.17
9031	2.78	2.67	3.56	3.11	3.00	2.89	3.00
9616	4.44	4.56	4.67	4.44	4.33	4.56	4.50
average	3.43	3.35	3.72	3.52	3.43	3.32	

Rater 4

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	4	2	3	2	3	3	2.83
0431	3	2	3	1	3	1	2.17
0684	5	5	5	5	5	5	5.00
1568	4	3	3	4	3	3	3.33
2094	3	4	5	5	5	5	4.50
3824	4	4	5	5	5	5	4.67
7375	4	4	4	5	5	4	4.33
7633	3	2	5	5	4	4	3.83
8473	4	3	3	3	4	4	3.50
9031	3	3	3	3	3	3	3.00
9616	5	5	5	5	5	5	5.00
average	3.82	3.36	4.00	3.91	4.09	3.82	

Average

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3.00	2.89	3.22	2.89	2.56	2.44	2.83
0431	2.56	2.67	3.00	2.89	2.44	2.33	2.65
0684	3.89	4.00	4.22	4.33	4.33	4.22	4.17
1568	3.22	3.11	3.33	2.78	3.00	2.89	3.06
2094	3.67	3.78	4.22	4.00	3.89	4.00	3.93
3824	4.00	3.56	4.00	3.89	3.56	3.44	3.74
7375	4.00	3.78	4.00	4.00	3.89	3.78	3.91
7633	3.00	2.89	3.33	3.33	3.33	3.00	3.15
8473	3.22	3.00	3.33	3.00	3.44	3.00	3.17
9031	2.78	2.67	3.56	3.11	3.00	2.89	3.00
9616	4.44	4.56	4.67	4.44	4.33	4.56	4.50
average	3.43	3.35	3.72	3.52	3.43	3.32	

Rater 5

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	4	4	4	3	2	2	3.17
0431	3	4	4	5	3	2	3.50
0684	5	5	5	3	4	2	4.00
1568	5	4	5	3	4	2	3.83
2094	3	4	5	4	4	4	4.00
3824	4	3	5	5	5	2	4.00
7375	5	3	4	5	5	2	4.00
7633	4	5	5	5	5	3	4.50
8473	3	3	5	3	5	2	3.50
9031	4	3	5	4	4	2	3.67
9616	5	5	5	5	5	4	4.83
average	4.10	3.91	4.73	4.09	4.18	2.45	

Average

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3.00	2.89	3.22	2.89	2.56	2.44	2.83
0431	2.56	2.67	3.00	2.89	2.44	2.33	2.65
0684	3.89	4.00	4.22	4.33	4.33	4.22	4.17
1568	3.22	3.11	3.33	2.78	3.00	2.89	3.06
2094	3.67	3.78	4.22	4.00	3.89	4.00	3.93
3824	4.00	3.56	4.00	3.89	3.56	3.44	3.74
7375	4.00	3.78	4.00	4.00	3.89	3.78	3.91
7633	3.00	2.89	3.33	3.33	3.33	3.00	3.15
8473	3.22	3.00	3.33	3.00	3.44	3.00	3.17
9031	2.78	2.67	3.56	3.11	3.00	2.89	3.00
9616	4.44	4.56	4.67	4.44	4.33	4.56	4.50
average	3.43	3.35	3.72	3.52	3.43	3.32	

Rater 6

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3	4	3	3	4	2	3.17
0431	2	2	3	3	2	2	2.33
0684	3	4	3	5	4	4	3.83
1568	3	4	3	3	3	4	3.33
2094	4	4	3	3	3	3	3.33
3824	4	4	3	3	4	3	3.50
7375	5	5	5	4	3	4	4.33
7633	2	2	2	2	3	2	2.17
8473	3	2	4	4	4	3	3.33
9031	3	2	3	3	3	3	2.83
9616	5	5	5	5	5	5	5.00
average	3.36	3.45	3.36	3.45	3.45	3.18	

Average

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3.00	2.89	3.22	2.89	2.56	2.44	2.83
0431	2.56	2.67	3.00	2.89	2.44	2.33	2.65
0684	3.89	4.00	4.22	4.33	4.33	4.22	4.17
1568	3.22	3.11	3.33	2.78	3.00	2.89	3.06
2094	3.67	3.78	4.22	4.00	3.89	4.00	3.93
3824	4.00	3.56	4.00	3.89	3.56	3.44	3.74
7375	4.00	3.78	4.00	4.00	3.89	3.78	3.91
7633	3.00	2.89	3.33	3.33	3.33	3.00	3.15
8473	3.22	3.00	3.33	3.00	3.44	3.00	3.17
9031	2.78	2.67	3.56	3.11	3.00	2.89	3.00
9616	4.44	4.56	4.67	4.44	4.33	4.56	4.50
average	3.43	3.35	3.72	3.52	3.43	3.32	

Rater 7

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3	4	3	3	1	2	2.67
0431	3	2	4	4	1	3	2.83
0684	3	4	5	4	4	5	4.17
1568	4	4	5	4	4	3	4.00
2094	4	5	5	5	5	4	4.67
3824	4	4	5	5	4	5	4.50
7375	4	3	4	4	4	5	4.00
7633	3	3	4	3	3	1	2.83
8473	3	4	3	3	3	2	3.00
9031	3	5	5	4	4	5	4.33
9616	5	5	5	4	4	5	4.67
average	3.55	3.91	4.36	3.91	3.36	3.64	

Average

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3.00	2.89	3.22	2.89	2.56	2.44	2.83
0431	2.56	2.67	3.00	2.89	2.44	2.33	2.65
0684	3.89	4.00	4.22	4.33	4.33	4.22	4.17
1568	3.22	3.11	3.33	2.78	3.00	2.89	3.06
2094	3.67	3.78	4.22	4.00	3.89	4.00	3.93
3824	4.00	3.56	4.00	3.89	3.56	3.44	3.74
7375	4.00	3.78	4.00	4.00	3.89	3.78	3.91
7633	3.00	2.89	3.33	3.33	3.33	3.00	3.15
8473	3.22	3.00	3.33	3.00	3.44	3.00	3.17
9031	2.78	2.67	3.56	3.11	3.00	2.89	3.00
9616	4.44	4.56	4.67	4.44	4.33	4.56	4.50
average	3.43	3.35	3.72	3.52	3.43	3.32	

Rater 9

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3	1	3	3	2	3	2.50
0431	1	3	1	1	2	1	1.50
0684	3	4	4	4	4	4	3.83
1568	2	1	1	1	2	1	1.33
2094	4	4	4	4	3	4	3.83
3824	4	2	3	2	2	2	2.50
7375	3	3	3	3	3	3	3.00
7633	2	4	2	2	2	2	2.33
8473	4	4	3	2	3	2	3.00
9031	1	2	2	2	2	2	1.83
9616	3	4	4	4	4	4	3.83
average	2.73	2.91	2.73	2.55	2.4	2.55	

Average

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3.00	2.89	3.22	2.89	2.56	2.44	2.83
0431	2.56	2.67	3.00	2.89	2.44	2.33	2.65
0684	3.89	4.00	4.22	4.33	4.33	4.22	4.17
1568	3.22	3.11	3.33	2.78	3.00	2.89	3.06
2094	3.67	3.78	4.22	4.00	3.89	4.00	3.93
3824	4.00	3.56	4.00	3.89	3.56	3.44	3.74
7375	4.00	3.78	4.00	4.00	3.89	3.78	3.91
7633	3.00	2.89	3.33	3.33	3.33	3.00	3.15
8473	3.22	3.00	3.33	3.00	3.44	3.00	3.17
9031	2.78	2.67	3.56	3.11	3.00	2.89	3.00
9616	4.44	4.56	4.67	4.44	4.33	4.56	4.50
average	3.43	3.35	3.72	3.52	3.43	3.32	

Rater 10

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3	3	4	4	3	4	3.50
0431	4	4	4	4	4	4	4.00
0684	4	3	3	4	4	4	3.67
1568	4	4	5	4	3	4	4.00
2094	4	3	5	4	4	4	4.00
3824	4	4	4	5	4	3	4.00
7375	4	4	4	4	3	4	3.83
7633	3	3	4	3	2	3	3.00
8473	3	2	2	3	2	3	2.50
9031	4	2	4	4	3	3	3.33
9616	4	4	4	4	3	4	3.83
average	3.73	3.27	3.91	3.91	3.18	3.64	

Average

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3.00	2.89	3.22	2.89	2.56	2.44	2.83
0431	2.56	2.67	3.00	2.89	2.44	2.33	2.65
0684	3.89	4.00	4.22	4.33	4.33	4.22	4.17
1568	3.22	3.11	3.33	2.78	3.00	2.89	3.06
2094	3.67	3.78	4.22	4.00	3.89	4.00	3.93
3824	4.00	3.56	4.00	3.89	3.56	3.44	3.74
7375	4.00	3.78	4.00	4.00	3.89	3.78	3.91
7633	3.00	2.89	3.33	3.33	3.33	3.00	3.15
8473	3.22	3.00	3.33	3.00	3.44	3.00	3.17
9031	2.78	2.67	3.56	3.11	3.00	2.89	3.00
9616	4.44	4.56	4.67	4.44	4.33	4.56	4.50
average	3.43	3.35	3.72	3.52	3.43	3.32	

Rater 11

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	2	2	3	2	2	2	2.17
0431	2	2	2	2	2	2	2.00
0684	4	4	5	5	5	5	4.67
1568	2	2	2	1	2	2	1.83
2094	4	4	3	3	3	3	3.33
3824	5	4	5	3	1	3	3.50
7375	4	4	4	4	4	4	4.00
7633	4	3	3	3	3	4	3.33
8473	3	3	3	3	4	4	3.33
9031	2	2	3	2	2	2	2.17
9616	5	5	5	5	4	5	4.83
average	3.36	3.18	3.45	3.00	2.901	3.23	

Average

	vowels	consonants	word stress	sentence stress	intonation	rhythm	average
0078	3.00	2.89	3.22	2.89	2.56	2.44	2.83
0431	2.56	2.67	3.00	2.89	2.44	2.33	2.65
0684	3.89	4.00	4.22	4.33	4.33	4.22	4.17
1568	3.22	3.11	3.33	2.78	3.00	2.89	3.06
2094	3.67	3.78	4.22	4.00	3.89	4.00	3.93
3824	4.00	3.56	4.00	3.89	3.56	3.44	3.74
7375	4.00	3.78	4.00	4.00	3.89	3.78	3.91
7633	3.00	2.89	3.33	3.33	3.33	3.00	3.15
8473	3.22	3.00	3.33	3.00	3.44	3.00	3.17
9031	2.78	2.67	3.56	3.11	3.00	2.89	3.00
9616	4.44	4.56	4.67	4.44	4.33	4.56	4.50
average	3.43	3.35	3.72	3.52	3.43	3.32	