Doctoral Dissertation

Linguistic Influences on Comprehensibility in Japanese EFL Learners' Speech

Faculty of Humanities,
Graduate School of Humanities,
Nagoya University

Ryosuke Mikami

March 2024

Abstract

In the field of second language (L2) speaking research, the concept of comprehensibility—broadly defined as the degree of ease to which listeners understand the speech of L2 speakers—has gained increasing attention from researchers and educators as a desirable learning goal for L2 learners. Researchers have examined a range of linguistic features that influence judgments of comprehensibility, aiming to enhance it effectively. They have consistently found that comprehensibility is affected by various linguistic factors—including pronunciation, fluency, vocabulary, grammar, and discourse. However, the relative contributions of these factors to comprehensibility remains inconclusive, as they vary based on numerous non-linguistic factors inherent in speakers and listeners. This study aims to address this research gap by exploring the potential impact of an additional factor—the subjectivity of linguistic assessments—on the linguistic features contributing to comprehensibility. To this end, a series of analyses was conducted. Correlation, principal component, and hierarchical multiple regression analyses consistently indicated that speech features such as pronunciation accuracy and speech rate exerted a more substantial influence on comprehensibility than lexicogrammatical accuracy and complexity. This observation held true regardless of the specific types of linguistic assessments employed. Furthermore, multivariate

analyses of variance revealed that the extent of influence of these linguistic features on comprehensibility varied depending on the comprehensibility levels. Specifically, for speakers with low to intermediate comprehensibility levels, pronunciation accuracy was more critical than lexicogrammatical accuracy and sophistication. Additionally, relatively smaller pronunciation units—such as segmental and word stress accuracy—had a more pronounced impact on comprehensibility than relatively larger units—such as rhythm and intonation accuracy. For speakers with intermediate to high comprehensibility levels, both pronunciation accuracy and fluency assumed equal importance alongside lexicogrammatical accuracy and sophistication. Moreover, in contrast to speakers with lower comprehensibility levels, relatively larger units of pronunciation—such as rhythm and intonation accuracy—were more pivotal in judgments of comprehensibility than relatively smaller units, such as segmental and word stress accuracy. In light of these findings, this study explores the assessment of comprehensibility from the perspective of raters' listening processes and provides relevant pedagogical implications.

Acknowledgments

I wish to express my profound gratitude to those who have provided unwavering support throughout my extensive Ph.D. journey. First and foremost, I would like to convey my sincere appreciation to my academic mentor, Professor Junko Yamashita. She has generously offered numerous opportunities for me to discuss the progress of my research project and has provided a wealth of constructive feedback. Additionally, I am grateful to Professor Masatoshi Sugiura and Professor Remi Murao for their invaluable guidance during our seminar classes, which has significantly enriched my academic pursuits.

Second, I would like to extend my thanks to my colleagues. They have played a pivotal role in helping me formulate a research plan, gather and analyze data, write my dissertation, and have provided ample opportunities for critical discussions and refining my research. Their unwavering encouragement and support, particularly during moments of personal discouragement, have been immensely motivating.

Third, I express my gratitude to all the participants who generously contributed their research data, without which my work would not have been possible.

Lastly, I must acknowledge my family's unwavering support and understanding throughout this journey. Their consistent support has been the cornerstone of my success.

I recognize that this achievement would have been unattainable without the steadfast support of these exceptional individuals. I extend my heartfelt appreciation to all of them for their contributions to my Ph.D. program.

Ryosuke Mikami

Author, Faculty of Humanities

Graduate School of Humanities, Nagoya University

Contents

List of Tables

List of Figures

List of Appendixes

Chapter 1 Introduction

In the field of learning second language (L2) speaking, there is a consensus among researchers and practitioners regarding the primary focus of adult L2 learners: Adult L2 learners should prioritize achieving comprehensibility (broadly defined as the degree of ease/difficulty with which interlocutors understand L2 speech) over pursuing native-like proficiency, considering attainability and realistic goals for L2 learners. This necessity comes from several factors, including maturational learning constraints (Abrahamsson, 2012; Abrahamsson & Hyltenstam; Flege et al., 1995; Munro & Mann; 2005), the significance of communicative effectiveness (Ludwig & Mora, 2017; Munro & Derwing, 1995b) and the practical importance within language testing contexts (Educational Testing Service, 2023).

Researchers are committed to investigating how L2 learners can effectively enhance the comprehensibility of their spoken L2 with the aim of establishing effective language programs and syllabi. An expanding body of research has explored the linguistic features related to L2 comprehensibility (e.g., Isaacs & Trofimovich, 2012; Saito et al., 2016; 2017). These endeavors have consistently unveiled the multifaceted nature of comprehensibility, encompassing a broad spectrum of linguistic features, including pronunciation, fluency, vocabulary, grammar, and discourse. Nonetheless, the

precise relationships between these features and their respective contributions to comprehensibility remain inconclusive due to variations arising from additional non-linguistic factors.

The present study aims to bridge this research gap by investigating the potential influence of an additional non-linguistic factor—the subjectivity inherent in linguistic assessments—on L2 comprehensibility. This dissertation has been structured as follows to achieve this objective.

Chapter 2 explores the background information essential for the study. It begins by providing a more precise definition of comprehensibility. Subsequently, the question of why L2 learners should prioritize comprehensibility is explored. This discussion is grounded in its relevance to language acquisition, its impact on effective communication, and its significance within the context of language testing. Following this, previous research investigating the connection between comprehensibility and various linguistic features is reviewed, along with their respective contributions to comprehensibility. Additionally, the role of factors that can moderate the relationship between these linguistic features and comprehensibility is examined. Lastly, this section identifies unresolved issues concerning the linguistic correlates of comprehensibility and outlines the direction of this study.

In Chapter 3, the experiment conducted in this study is detailed, beginning with an outline of the research objectives and two research questions for investigation. Subsequently, an in-depth description of the data collection methodology is provided, divided into three components: the collection of speech samples, the assessment of comprehensibility of L2 speech, and the linguistic evaluation of the collected tokens.

Chapter 4 encompasses the data analysis procedures and their corresponding results. In this section, five primary statistical analyses are conducted to elucidate the impact of various linguistic features on comprehensibility. First, descriptive statistics are computed, and data distributions are examined for comprehensibility ratings and linguistic features. Second, correlation analyses are employed to ascertain which specific linguistic features within the domains of pronunciation, fluency, lexis, and grammar are associated with comprehensibility judgments. Third, principal component analyses are conducted to reduce the number of linguistic variables into smaller number and make the interpretation easier. Fourth, hierarchical multiple regression analyses are utilized to gauge the overall influence of these linguistic features on comprehensibility. Finally, multivariate analyses of variance are executed to explore whether the linguistic influence on comprehensibility varies across differing comprehensibility levels.

Chapter 5 provides a discussion of the present findings: the relationship between linguistic features and comprehensibility. This study draws on these discussions to derive pedagogical implications for L2 speakers across various proficiency levels.

Lastly, Chapter 6 provides a summary of the study and concludes by outlining potential future avenues for research within the domain of comprehensibility studies in L2 speaking.

Chapter 2 Background

2.1 Definition of Comprehensibility

It is essential to understand the distinction among three major constructs: *accentedness*, *intelligibility*, and *comprehensibility*. These constructs are traditionally discussed in the field of L2 pronunciation. Levis (2005) introduces two goals for L2 learners: the *nativeness* and *intelligibility principles*. The former straightforwardly prioritizes reducing foreign accents and sounding "native-like." In recent studies, this term has been interchangeably referred to as *accentedness*, defined as the degree to which one's pronunciation deviates from the "native norm" (Derwing & Munro, 2009). This construct is frequently operationalized through subjective scaler ratings such as Likert-type (e.g., Derwing & Munro, 1997) and 1000-point sliding scales (e.g., Saito et al., 2017).

In contrast, the *intelligibility principle* compromises accent reduction and pursues being understandable for the interlocutors. When discussing intelligibility, care must be taken to interpret the term and discriminate between its narrow and broad senses. In a narrow sense, intelligibility encompasses the degree to which L2 speech is clear and perceivable to the interlocutors (Derwing & Munro, 2009). It is often measured as the

number of words the listener could transcribe orthographically divided by the total number of words produced.

In contrast, intelligibility in a broad sense has been called *comprehensibility* in recent literature (e.g., Isaacs and Trofimovich, 2012). From this point forward, this study uses the term "comprehensibility" to denote intelligibility in this sense. Comprehensibility refers to the extent of ease/difficulty to which the listeners need to reconstruct the overall message or content of speech conveyed by L2 speaker (Derwing & Munro, 2009). In other words, it denotes the perceived amount of effort the listeners need for reconstructing the meaning of the uttered L2 speech. This construct is frequently operationalized using Likert-type and 1000-point sliding scales.

Notably, intelligibility and comprehensibility are related but separate constructs in two aspects. The first is associated with the degree of linguistic processing. Because intelligibility is operationalized as the number of words the listener could recognize, it primarily focuses on surface-level aspects of pronunciation, such as the accurate articulation of individual words and the appropriate use of stress and intonation patterns. In contrast, as comprehensibility pertains to the construction of meaning, it encompasses a broader spectrum of linguistic features beyond pronunciation and temporal features, including lexical and grammatical accuracy and sophistication, as

well as discourse structure (Crowther, Trofimovich, Isaacs, & Saito, 2015; Isaacs & Trofimovich; Saito et al., 2016; 2017).

The second aspect concerns the idea that highly intelligible speech is not necessarily highly comprehensible speech. Specifically, L2 speech with low comprehensibility (i.e., much listening effort is needed to reconstruct the meaning) can still be highly intelligible (i.e., many words can be recognized) for the interlocutors (Munro & Derwing, 1995a).

In recent studies, comprehensibility has gained considerable attention as a key aim in achieving successful oral communication for L2 learners (e.g., Trofimovich & Isaacs, Saito et al., 2016; 2017). Therefore, the current study centers on comprehensibility as an ideal learning goal.

2.2 Why Target Comprehensibility?

2.2.1 Difficulty in Achieving "Native-like" Proficiency Levels in Various Linguistic Domains

The principal rationale for striving for comprehensibility is that most L2 learners cannot reach native-like proficiency in speaking, even with extensive learning or exposure to native input in the target language environment.

Numerous scholars have reached a consensus about the multifaceted nature of speaking proficiency, encompassing intricate subdimensions (e.g., Iwashita et al., 2008; Michel, 2017). For instance, Iwashita et al. (2008) categorized speaking proficiency into subcomponents such as pronunciation, fluency, lexis, and grammar. As some L2 learners aspire to attain native-like speaking proficiency in such domains (Derwing, 2003), scholars have, over time, investigated L2 learners' potential for achieving native-like L2 proficiency in diverse linguistic domains (e.g., Abrahamsson, 2012; Abrahamsson & Hyltenstam, 2009; Birdsong, 2006; Meara & Bell, 2001; Mora & Valls-Ferrer, 2012; Pang & Skehan, 2014; Saito & Saito, 2017; Trofimovich & Baker, 2006). Regrettably, a substantial body of research indicated that a limited number of L2 learners attain native-like proficiency.

For example, Flege et al. (1995) examined the influence of the age of learning (AOL) on perceived foreign accents among Italian learners of English who began residing in Canada between the ages of 2–23. They included native English speakers as a control group. All participants were required to produce English sentences, which were assessed by native English listeners for perceived foreign accents using a continuous scale. The findings from a regression analysis revealed a significant impact of AOL on perceived foreign accents, with approximately 60% of variance explained.

Specifically, learners who arrived in Canada at older ages exhibited stronger foreign accents, while native speakers exhibited the least accents. Interestingly, the length of residence in Canada had little effect on foreign accents, with less than 2% of the variance explained. This observation implied that even with extensive exposure to native input, L2 learners struggled to develop their pronunciation proficiency.

Munro and Mann (2005) also explored the impact of age of immersion (AOI) on the degree of perceived accent (DPA) among Mandarin speakers of English who initiated their English learning in the United States between the ages of 3–16. The speakers were asked to produce English expressions, including individual words, sentences, a paragraph, and spontaneous speech. Subsequently, native English listeners evaluated these utterances for DPA on a continuous scale. A linear regression analysis revealed a significant negative association between AOI and DPA. This finding suggested that individuals who began their English learning later exhibited stronger foreign accents. It highlighted the challenge L2 learners face in achieving native-like pronunciation, particularly when they initiated their foreign language learning later. Furthermore, time spent residing in the United States (an average of 11.6 years) suggested that attaining native-like pronunciation is challenging for L2 learners despite their extensive exposure to the target language.

Trofimovich and Baker (2006) demonstrated adult L2 learners' difficulty attaining native-like suprasegmental accuracy and fluency levels. The researchers compared suprasegmental accuracy and fluency among Korean learners of English who had resided in the United States for between three months and ten years compared to native English speakers. The findings revealed that even those L2 learners with a decade of experience in the target language environment spoke at a slower speech rate and exhibited lower accuracy in suprasegmental aspects—stress, rhythm, and intonation—compared to native speakers.

Abrahamsson (2012) also explored the impact of the age of onset (AO) on L2 proficiency, focusing on perception-related grammatical and phonetic skills. The study targeted Spanish learners of Swedish who had initiated their residence in Sweden between the ages of 1–30. Native Swedish speakers were included as a control group. The L2 learners were categorized into *early* and *late* groups based on their AO. The *early* group comprised L2 learners who commenced residency between the ages of 1–15. In contrast, the *late* group consisted of those who initiated residence between the ages of 16–30. All participants underwent two tests: an auditory grammaticality judgment test and a phonetic perception test. In the grammaticality judgment test, participants were asked to evaluate the grammatical correctness of Swedish sentences,

each containing one of four morphosyntactic structures or features from Swedish grammar that pose particular challenges for L2 learners. In the phonetic perception test, participants were asked to discern whether the presented stop consonants were voiced or voiceless. The included items in this test were recognized as problematic for L2 learners in distinguishing between voiced and voiceless sounds. ANOVAs were conducted to compare the participants' performance on both tests. These analyses revealed that the *early* group performed less proficiently than the native group, while the *late* group performed even less proficiently than the *early* group on both tests. Additionally, correlation analyses were conducted independently for the *early* and *late* groups to explore the relationship between test performance and AO. These analyses demonstrated that the *early* group displayed significant correlations between test performance and AO, whereas the *late* group did not exhibit significant correlations. These findings suggested that even L2 learners who commence exposure to an L2 environment earlier than the mid-teen years struggled to attain native-like grammatical and phonetic intuitions. Additionally, L2 learners who entered an L2 environment after their mid-teens showed limited improvement in their grammatical and phonetic perceptions.

Abrahamsson and Hyltenstam (2009) examined the influence of age of onset (AO) on various aspects of linguistic proficiency. This research is relevant because it comprehensively covered linguistic features. The study consists of two distinct parts. In Part 1, the focus was on 195 highly advanced Spanish learners of Swedish, with a range of AO spanning less than 1–47 years. Thirty native Swedish speakers were also included. Participants were asked to produce spontaneous speech, and their speech samples were evaluated for native-likeness by native judges. Consequently, 41 participants who exhibited performance levels within the range of native speakers and met most background criteria (i.e., age, sex, frequency of daily L1 use) were selected for further linguistic scrutiny in Part 2. Part 2 included these 41 learners and additional 15 native speakers. This phase involved the administration of a battery of 10 language assessments, focusing on speech production and perception, grammar, vocabulary, pragmatics, idiomatic expressions, and proverbs. The tests and tasks were designed to be complex to induce a substantial degree of difficulty and cognitive load, even for native speakers. The analysis revealed that only three of the 41 learners achieved proficiency levels within the range of native speakers across all 10 measures. Remarkably, these learners had AO at 3, 7, and 8 years. The researchers concluded that

achieving native-like L2 proficiency remains, in principle, unattainable for adult learners.

In the context of these challenges, numerous researchers and practitioners concur that comprehensibility should replace the unattainable pursuit of native-like proficiency to enable L2 learners to communicate effectively (e.g., Levis, 2005; Munro & Derwing, 1999).

2.2.2 The Relation of Comprehensibility to Rate of Speech Understanding

Comprehensibility is important for authentic communication because less comprehensible speech can hinder effective communication. Research has shown that native and non-native listeners require more time to comprehend L2 speech that they perceive as less comprehensible (Ludwig & Mora, 2017; Munro & Derwing, 1995b).

For example, Munro and Derwing (1995b) investigated the impact of comprehensibility on the time necessary for understanding L2 speech. Mandarin speakers of English recorded English sentences, the truth value of which could easily be determined by native listeners based on everyday knowledge (e.g., “Elephants are big animals” and “Most people wear hats on their feet”). Native English listeners subsequently listened to these sentences and verified their truth value as quickly as possible. Additionally, these listeners assessed the comprehensibility of the sentences.

The findings indicated that L2 speech characterized by low comprehensibility took longer for native listeners to understand than L2 speech with moderate to high comprehensibility.

Ludwig and Mora (2017) also explored the relationship between comprehensibility and the speed of understanding L2 speech for L2 and native listeners. The researchers gathered L2 English sentences read aloud by Catalan and German learners of English. Subsequently, different groups of Catalan and German learners and L1 English speakers engaged in a sentence verification task, which involved listening to the speech samples and assessing the truth value of the sentences following Munro and Derwing's method (1995b). In parallel, the participants evaluated the comprehensibility of the sentences. The findings demonstrated a significant correlation between comprehensibility and the speed of understanding L2 speech for both native and non-native listeners. This suggested that both groups of listeners require more time to understand L2 speech with lower comprehensibility.

As reviewed above, the comprehensibility of speech is closely linked to the speed at which both native and non-native listeners understand L2 speech. More specifically, less comprehensible speech has the potential to impede the smooth flow of communication.

2.2.3 Integration into Speaking Rating Criteria in High-stakes Language Tests

The concept of comprehensibility has recently gained significance in the assessment criteria for high-stakes language tests, such as TOEFL iBT and IELTS. To illustrate, the TOEFL iBT speaking section comprises both Independent Speaking Tasks and Integrated Speaking Tasks, each assessed based on three key components: Delivery, Language Use, and Topic Development. *Delivery* pertains to sound aspects like pronunciation and fluency. *Language Use* evaluates how effectively test takers employ grammar and vocabulary to articulate their ideas. *Topic Development* examines how coherently they construct their responses by linking thoughts and elaborating on ideas. Among these components, comprehensibility frequently surfaces within the *Delivery* category. Although the term "comprehensibility" may not explicitly be used, it is often expressed as "listener effort." In Independent Speaking Rubrics, for instance, a score of three out of a maximum of four is characterized as: "generally clear, with some fluidity of expression, though minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times (though overall intelligibility is not significantly affected)" (Educational Testing Service, 2023, Section for a Score of three in the Delivery category). Similarly, a score of two is characterized as: "basically

intelligible, though listener effort is needed" (Educational Testing Service, 2023, Section for a Score of 2 in the Delivery category).

In Integrated Speaking Rubrics, the score of three, for example, is characterized as follows: "[speech] may require some listener effort at times. Overall intelligibility remains good, however." (ETS, 2023, Section for a Score of four in the Language Use category). On rarer occasions, comprehensibility is a consideration in the evaluation criteria for *Language Use*, which focuses on grammar and vocabulary use. For example, a maximum score of four is characterized as: "Though some minor (or systematic) errors or imprecise use may be noticeable, they do not require listener effort (or obscure meaning)" (Educational Testing Service, 2023, Section for a Score of four in the Language Use category).

Comprehensibility is also a critical assessment criterion within the IELTS *Speaking Band Descriptors*. Candidates' performance is assessed across four distinct categories, each spanning nine bands: *Fluency and coherence, Lexical resource, Grammatical range and accuracy,* and *Pronunciation*. Comprehensibility features predominantly within the Pronunciation category. International Development Program (IDP) Education (2023) explicitly outlines this aspect. In particular, this category assesses "How easy it is to understand what you say?" (IDP Education, 2023, 4.

Pronunciation, “What do the band descriptors mean?”). For instance, a speech receiving a maximum score of nine is described as: “Can be effortlessly understood throughout” (IDP Education, 2023, p. 1). Likewise, a speech achieving a score of six is characterized as: “Can generally be understood throughout without much effort” (IDP Education, 2023, p. 1). Similarly, comprehensibility is detailed within the band descriptors for other score levels.

Above all, comprehensibility plays a role in the evaluation criteria of high-stakes language tests such as TOEFL iBT and IELTS. Consequently, L2 learners should endeavor to make themselves understood and ensure that the interlocutors can *easily* understand their message.

In summary, three key factors discussed from 2.2.1 to 2.2.3 suggest that we should consider comprehensibility as an attainable, realistic, and important goal for L2 speaking.

2.3 Linguistic Correlates of Comprehensibility

Comprehensibility and its linguistic correlates are long-standing research topics. The seminal research by Isaacs and Trofimovich (2012) represents a watershed. In earlier studies, comprehensibility was often considered a singular aspect of pronunciation proficiency (Derwing & Munro, 1997; Munro & Derwing, 1995; Hahn,

2004). Consequently, earlier investigations into the link between comprehensibility and linguistic features primarily focused on aspects of sound features, such as pronunciation and fluency. Occasionally, they also included grammatical and lexical features. These early studies collectively indicate links between comprehensibility and various linguistic features. However, a notable limitation is the inability to determine the relative importance of these linguistic features to comprehensibility due to the omission of effect size computations such as correlation coefficients.

Isaacs and Trofimovich (2012) broke this tradition by adopting a more comprehensive view, redefining comprehensibility as a multifaceted speaking construct. This expanded perspective encompassed pronunciation and fluency and integrated considerations of lexical and grammatical sophistication, and discourse features. Subsequently, researchers began exploring the relative contributions of these linguistic domains to comprehensibility—pronunciation, fluency, lexis, grammar, and discourse.

Munro and Derwing (1995) research is one of the earliest studies to investigate the connection between comprehensibility and various linguistic features, including pronunciation and grammatical accuracy. They gathered brief spontaneous speech samples from advanced Mandarin speakers of English. Subsequently, native English speakers evaluated their comprehensibility on a 9-point scale. Furthermore, the

researchers assessed the speech samples for phonemic errors, phonetic errors, overall intonation goodness, and grammatical errors. Correlation analyses between comprehensibility and each linguistic feature were repetitively conducted for each listener, followed by an analysis of the number of listeners who exhibited a significant correlation. The findings indicated that phonetic and phonemic errors were correlated with comprehensibility in 11% and 44% of the listeners, respectively. Moreover, grammatical errors and intonation goodness exhibited a significant correlation with comprehensibility for 56% and 83% of the listeners, respectively.

Derwing and Munro (1997) replicated these findings among L2 English speakers with intermediate proficiency levels, encompassing a range of L1 backgrounds—Cantonese, Japanese, Polish, and Spanish. They also noted that speaking rate was associated to comprehensibility. Furthermore, Hahn (2004) reported sentence stress errors were related to comprehensibility.

Previous research on how lexical features affect listeners' perceptions of L2 speech is limited. While Fayer and Krasisnski (1987) is frequently cited as investigating the impact of lexical features on comprehensibility, it primarily focused on intelligibility rather than comprehensibility. Also, even though their study assessed the frequency of incorrect word choices, they combined it with other linguistic scores

(pronunciation and grammatical accuracy) and calculated mean linguistic scores to analyze the overall influence of these linguistic features on intelligibility. Thus, the specific impact of lexical features on intelligibility was not examined.

Trofimovich and Isaacs (2012) provide a seminal contribution in two key aspects. First, they extended their linguistic analysis beyond accuracy, encompassing factors such as lexical diversity and sophistication, and discourse structure. The second is the quantitative analysis of the relative importance of the various linguistic features. They achieve this by calculating correlation and regression coefficients. This research marks a significant turning point, as subsequent studies began incorporating a wider range of linguistic features and assessing their relative importance to comprehensibility.

Their study examined the impact of 19 linguistic features—including pronunciation, fluency, lexis, grammar, and discourse—on comprehensibility. The study involved 40 French learners who produced L2 English speech. The comprehensibility of the speech was assessed by 60 English L1 speakers. The findings revealed significant correlations between 18 of the features except pitch range and comprehensibility. In addition, multiple regression analysis was used to identify the most influential predictors of comprehensibility. Type frequency (i.e., a feature of

lexical diversity) was the strongest predictor, accounting for 64% of the variance, followed by word stress accuracy (16%) and grammatical accuracy (6%).

Saito et al. (2017) replicated Trofimovich and Isaacs (2012) using the same speech samples obtained from 40 French speakers of English. However, Saito et al. differed from the previous study by having 20 English L1 speakers use a 1000-point sliding scale to intuitively evaluate 11 linguistic features, encompassing pronunciation, fluency, lexis, grammar, and discourse. The 11 linguistic variables were subjected to principal component analysis to reduce the number of variables, resulting in two composite variables—pronunciation and lexicogrammar. They conducted a stepwise multiple regression analysis to assess the effect of these composite variables on comprehensibility, with comprehensibility as the dependent variable. The results indicated that pronunciation accounted for 50% of the variance in comprehensibility, while lexicogrammar explained 40%. In addition, five pronunciation and fluency features were similarly correlated with comprehensibility, as were five lexis, grammar, and discourse features (except discourse cohesion). These findings confirmed the importance of both pronunciation and lexicogrammar in comprehensibility assessments, which was consistent with Trofimovich and Isaacs (2012).

Saito et al. (2015) thoroughly examined lexical features in L2 English speech of French learners and their effects on comprehensibility, building upon Trofimovich and Isaacs (2012). The study analyzed 12 lexical variables across six lexical subcategories, including appropriateness, fluency, variation, sophistication, abstractness, and sense relations. Additionally, four pronunciation features—segmental errors, syllable structure errors, word stress errors, and intonation errors—were analyzed and combined into a composite pronunciation variable. Partial correlation analyses were employed to identify the specific contributions of lexical variables to listener evaluations of comprehensibility while controlling for the effects of pronunciation. The results indicated that eight features in lexical sub-domains of appropriateness, fluency, variation, abstractness, and sense relations were significantly associated with comprehensibility, with effect sizes ranging from small to large. Sophistication features did not display significant correlations with comprehensibility. Furthermore, the correlation between pronunciation and comprehensibility was moderate when relevant lexical variables were partialled out.

Lastly, Suzuki and Kormos (2020) also examined various linguistic correlates of comprehensibility. The participants were 40 Japanese learners of English who provided spontaneous speech samples. Subsequently, 10 L1 English speakers evaluated these

speech samples for their comprehensibility using a 9-point scale. Moreover, the researchers used automated linguistic assessment software to objectively analyze the speech samples for 23 linguistic features, encompassing the domains of pronunciation, fluency, lexis, grammar, and discourse. Correlation analyses indicated that 15 of the features from all the domains significantly correlated with comprehensibility, with correlations ranging from moderate to high. Furthermore, multiple regression analysis revealed that five features accounted for 92.1% in the total variance of the comprehensibility ratings. Among these features, articulation rate had the highest impact, accounting for 67.0% of the variance, followed by mid-clause pause duration (14.0%), morphological error ratio (5.7%), syllable structure error ratio (4.2%), and mid-clause pause ratio (0.9%).

In summary, comprehensibility is associated with various linguistic features, including pronunciation, fluency, vocabulary, grammar, and discourse. Overall, speech features, such as pronunciation and fluency, demonstrate stronger connections compared to lexical, grammatical, and discourse features (Crowther, Trofimovich, Isaacs, & Saito, 2015; Saito & Shintani, 2015; Saito et al., 2016; 2017; Suzuki & Kormos, 2020). However, the specific impact of these features can vary across studies

(e.g., syllable structure errors as observed in Suzuki & Kormos, 2020 and Trofimovich & Isaacs, 2012).

2.4 Additional Factors Affecting Linguistic Correlates of Comprehensibility

Expanding upon the findings discussed in the previous section, researchers further explored factors influencing the association between comprehensibility and linguistic features. The factors under consideration can be broadly classified into two categories: those related to the speaker and those related to the listener. The speaker-related factors encompass speakers' L1 background (Crowther, Trofimovich, Saito, and Isaacs, 2015), comprehensibility levels (Saito et al., 2016), and the impact of task design on eliciting L2 speech (Crowther, Trofimovich, Isaacs, and Saito, 2015). The listener-related factors include listeners' L1 background (Saito & Shintani, 2015, comparing monolingual L1 listeners with bilingual L1 listeners; Mikami, 2019, examining L2 listeners with diverse L1 backgrounds). The following paragraphs provide an overview of how these factors influence various linguistic correlates of comprehensibility.

The first speaker-related factor pertains to their L1 backgrounds. Crowther, Trofimovich, Saito, and Isaacs (2015) investigated how a speaker's L1 background influenced the linguistic features of comprehensibility. They collected spontaneous L2 speech samples using a picture narrative task from 45 L2 English learners whose L1s

were either Mandarin Chinese, Farsi, or Hindi-Urdu. Subsequently, 10 L1 English expert raters assessed these speech samples for comprehensibility and 10 linguistic features across the domains of pronunciation, fluency, vocabulary, grammar, and discourse, using 1000-point sliding scales. Principal component analysis was employed to cluster these 10 linguistic features into fewer variables, yielding two composite variables—pronunciation and lexicogrammar. A multiple regression analysis that encompassed all three speaker groups, with the composite scores as independent variables, found that lexicogrammar and pronunciation explained 49% and 21% of the variance, respectively. Furthermore, separate correlation analyses for each L1 group showed that only pronunciation had a significant correlation with comprehensibility for the Mandarin group, whereas only lexicogrammar demonstrated a significant correlation for the Hindi-Urdu group. Notably, neither pronunciation nor lexicogrammar exhibited significant correlations for the Farsi group. These findings suggested that a speaker's L1 background plays a crucial role in determining the linguistic features that impact comprehensibility.

Another factor relevant to speakers is the degree of comprehensibility. Saito et al. (2016) collected speech samples from 120 Japanese learners of English who completed three picture description tasks. Five L1 English speakers rated their comprehensibility

on a 9-point scale. Furthermore, five experienced L1 English speakers used 1000-point sliding scales to evaluate the samples for eight linguistic features, spanning the domains of pronunciation, fluency, lexis, and grammar. Correlation analyses revealed that six of these features were significantly correlated with comprehensibility, with effect sizes ranging from small to large. Principal component analysis was then conducted to cluster the eight features into a smaller number of variables, resulting in three composite scores: pronunciation, lexicogrammar accuracy, and lexicogrammar sophistication. The researchers then performed a stepwise multiple regression analysis, with comprehensibility as the dependent variable and the composite scores as the independent variables. The results demonstrated that pronunciation had the greatest contribution to comprehensibility, explaining 50% of the variance. Lexicogrammar accuracy and sophistication followed with explanatory power of 22% and 8%, respectively. Finally, to investigate the linguistic correlates of comprehensibility across different comprehensibility levels, the speakers were categorized into four groups based on their comprehensibility ratings: low beginners, high beginners, intermediate, and advanced speakers. An analysis of variance was employed to determine which linguistic features distinguished between these comprehensibility levels. The findings revealed that word stress and intonation accuracy effectively differentiated all four levels. In

contrast, speech rate and lexical appropriateness were only successful in distinguishing between the low-beginner and high-beginner groups. Additionally, grammatical accuracy proved to be a discriminative factor for the intermediate and advanced groups, as well as for the low- and high-beginner groups. These results suggested that the linguistic components contributing to comprehensibility vary depending on the level of comprehensibility.

The final factor related to speakers involves the influence of task design in eliciting L2 speech. Crowther, Trofimovich, Isaacs, and Saito (2015) explored whether the cognitive demands of different speaking tasks influence the relationship between comprehensibility and various linguistic features in L2 speech. They gathered spontaneous speech samples from 60 L2 English learners (grouped into four according to L1s: Mandarin Chinese, Hindi-Urdu, Farsi, and Romance languages). Two tasks were utilized to compare the effects of different tasks: the IELTS long-turn speaking task and the TOEFL iBT integrated speaking task. Each task imposed varying cognitive demands on the speakers, with the TOEFL task being more cognitively demanding because the TOEFL task involved handling unfamiliar factual information, necessitating reasoning and perspective-taking, which was not required in the IELTS task. Subsequently, 10 L1 English expert raters evaluated the speech samples for comprehensibility and 10

linguistic features. These features were then grouped into two composite variables of pronunciation and lexicogrammar using principal component analyses.

Two sets of multiple regression analyses were carried out for each test condition to investigate the overall impact of the composite variables on comprehensibility, encompassing all four speaker groups. The results revealed similar findings for both the IELTS and TOEFL tests. In the case of IELTS, pronunciation and lexicogrammar accounted for 60% and 14% of the variance, respectively. Similarly, for the TOEFL, pronunciation and lexicogrammar accounted for 71% and 17% of the variance, respectively. Furthermore, correlation analyses were performed to examine the influence of each linguistic feature on comprehensibility, depending on the test conditions across the speakers' L1 backgrounds. These analyses indicated a clear impact of task variation. In the IELTS task, comprehensibility was primarily associated with pronunciation and fluency features for three of the four groups. Only the Farsi group demonstrated associations with vocabulary, grammar, and discourse. In contrast, in the more cognitively demanding TOEFL task, comprehensibility was also related to linguistic features at the level of grammar, vocabulary, and discourse for all groups.

Listener-related factors were also investigated, with a focus on the language backgrounds of listeners. The first study is by Saito and Shintani (2015), investigating

how the language background of L1 listeners influences the comprehensibility of L2 speech. Specifically, they compared assessments of comprehensibility made by monolingual Canadian and bilingual Singaporean L1 English speakers. They gathered spontaneous speech samples from 50 Japanese learners of English using a timed-picture narrative task. Subsequently, Canadian and Singaporean evaluators rated the comprehensibility of these samples using a 9-point scale. Additionally, native English raters with experience in teaching English in ESL/EFL settings used 1000-point sliding scales to analyze the speech samples for eight linguistic features, encompassing features such as pronunciation, fluency, vocabulary, and grammar. The results of correlation analyses revealed that speech rate and accuracy in segments, prosody, and grammar were consistently linked to comprehensibility for both listener groups. Furthermore, the correlations between speech rate and accuracy in segments and prosody were more pronounced in Canadian listeners than in Singaporean listeners, while grammatical accuracy was equally associated with comprehensibility for both listener groups. Moreover, a correlation between lexical appropriateness and comprehensibility was found only in the Singaporean listeners. Furthermore, the results of multiple regression analyses showed that for the Canadian evaluators, the primary predictors of comprehensibility judgments were related to pronunciation and fluency (specifically,

segmentals and speech rate), explaining 79% of the variance. In contrast, grammatical accuracy played a minor role, explaining only 5% of the variance. Conversely, for the Singaporean evaluators, the findings demonstrated that pronunciation and speech rate accounted for 64% of the variance in comprehensibility judgments. Lexical features also contributed significantly, explaining 16% of the variance. These findings indicated that monolingual native raters primarily consider phonological accuracy and temporal features when making comprehensibility judgments, while bilingual native raters based their assessments on a wider range of linguistic information, including pronunciation, fluency, vocabulary, and grammar.

Another study that examined the impact of listeners' language backgrounds is Mikami (2019). In contrast to Saito and Shintani (2015), this research compared the assessment of comprehensibility between two groups of L2 English listeners whose L1s were either Mandarin Chinese or Japanese. A baseline group of L1 English listeners was also included in the study. Spontaneous speech samples were collected from Japanese learners of English through a storytelling task. Subsequently, three groups of listeners evaluated these speech samples for comprehensibility using a 9-point scale. In parallel, expert L1 English teachers assessed nine linguistic features of the speech samples, employing 1000-point sliding scales and covering aspects such as

pronunciation, fluency, vocabulary, and grammar. Correlation analyses were conducted separately for each group of listeners. These analyses revealed that, for both English and Chinese listener groups, comprehensibility was associated with only five pronunciation and fluency features. In the case of the Japanese listener group, four additional lexical and grammatical features correlated with comprehensibility. Multiple regression analyses showed that intonation accuracy was the sole predictor of comprehensibility judgments in both English and Chinese listener groups. In contrast, rhythm accuracy and lexical richness jointly influenced comprehensibility in the Japanese listener group. These findings showed that L2 speakers with differing L1 backgrounds pay attention to different linguistic features when assessing comprehensibility.

2.5 Consistencies and Discrepancies in Prior Findings

This section summarizes the earlier research on the linguistic correlates of comprehensibility, focusing on the commonalities and variations.

First, a consistent pattern is the *overall* association between comprehensibility and a diverse array of linguistic features. These features encompass pronunciation, fluency, vocabulary, grammar, and discourse (e.g., Crowther, Trofimovich, Saito, & Isaacs, 2015; Trofimovich & Isaacs, 2012). Furthermore, many studies, though not all, consistently revealed that speech features such as pronunciation accuracy and temporal

features have a greater impact on comprehensibility than lexical, grammatical, and discoursal features (Crowther, Trofimovich, Isaacs, & Saito, 2015; Mikami, 2019; Saito et al., 2016; 2017; Saito & Shintani, 2015; Suzuki & Kormos, 2020).

However, when comparing the influence of *specific* linguistic features on comprehensibility across previous studies, some inconsistencies are encountered. For instance, in the examination of the impact of syllable structure errors on comprehensibility, Trofimovich and Isaacs (2012) reported a small correlation of −.37 with comprehensibility, whereas Suzuki and Kormos (2020) found a larger correlation of −.79. More strikingly, Saito et al. (2016) and Suzuki and Kormos (2020) both investigated the effect of lexical diversity on comprehensibility, but their results were contradictory. Saito et al. (2016) revealed a significant correlation of .40 between lexical diversity and comprehensibility, while Suzuki and Kormos (2020) did not find a significant correlation. Similar disparities were also apparent in other linguistic features, such as articulation rate and lexical accuracy (Trofimovich & Isaacs, 2012; Suzuki & Kormos, 2020). These discrepancies pose challenges from a pedagogical perspective because curriculum developers, textbook writers, and practitioners cannot make evidence-based decisions about which specific linguistic features to teach or to include in curricula or textbooks.

As previously discussed, the understanding of the linguistic correlates of comprehensibility is a complex phenomenon, as it varies depending on a range of factors, and the study designs differ. Nonetheless, it is crucial to explore additional potential factors contributing to the divergent outcomes noted.

The initial step in addressing this complexity is to compare prior studies that used comparable research designs but obtained different outcomes: Trofimovich and Isaacs (2012) and Saito et al. (2017). Both analyzed the same set of speech data, obtained comprehensibility ratings from L1 English speakers, and examined similar linguistic features in the realms of pronunciation, fluency, vocabulary, grammar, and discourse. The divergent outcomes between the studies and potential factors that contribute to this inconsistency are explored in the following section.

2.6 Possible Factors Moderating the Relationship Between Comprehensibility and Linguistic Features

Trofimovich and Isaacs (2012) found that lexical diversity is the most robust predictor of comprehensibility, with pronunciation playing a minor role, while Saito et al. (2017) reported pronunciation and fluency as the strongest predictor, with lexicogrammar also playing a substantial role. Both studies controlled for the previously mentioned factors associated with speakers and listeners that impact linguistic features

of comprehensibility. This suggests that there might be additional factors influencing the relationship between comprehensibility and linguistic features. Upon closer examination, the subjectivity in linguistic assessments arose as a potential confounding factor.

Trofimovich and Isaacs (2012) objectively assessed 19 linguistic features of L2 speech. In contrast, Saito et al. (2017) relied on subjective assessments, wherein L1 English speakers used 1000-point sliding scales to intuitively evaluate these aspects after listening to speech samples and reviewing their transcripts. This difference may influence the linguistic features that relate to comprehensibility.

In addition to exploring the linguistic features contributing to comprehensibility, Saito et al. (2017) had an additional objective—to verify the accuracy of subjective linguistic assessment utilizing a 1000-point sliding scale. The researchers re-analyzed the same speech data set employed in Trofimovich and Isaacs (2012), including pronunciation, fluency, vocabulary, grammar, and discourse, with acoustic analysis software or textual analyses. Native English speakers listened to the speech samples, reviewed their transcripts, and intuitively evaluated 11 linguistic features—spanning pronunciation, fluency, vocabulary, grammar, and discourse—which corresponded to the objective linguistic features employed in Trofimovich and Isaacs (2012). Saito et al.

used 1000-point sliding scales in the assessments. After that, they conducted correlation analyses between the objective linguistic features assessed by Trofimovich and Isaacs (2012) and their subjectively assessed linguistic features. In this analysis, 10 of the 11 subjectively rated features demonstrated significant correlations with their corresponding objectively rated linguistic features. These findings led the researchers to conclude that subjective linguistic assessments are a valid means of evaluating various linguistic features.

It is noteworthy, however, that the correlation coefficients varied widely, ranging from .43 to .79. Strikingly, discourse cohesion failed to demonstrate significant correlations with their corresponding objective counterparts. Consequently, the interchangeable use of subjective and objective linguistic assessments should be exercised cautiously. Ultimately, these disparities between subjective and objective linguistic assessments could explain the differing findings regarding the linguistic correlates of comprehensibility between Trofimovich and Isaacs (2012) and Saito et al. (2017).

Indeed, prior studies that utilized either subjective or objective linguistic assessments yielded differing results regarding the linguistic features that affect comprehensibility. Suzuki and Kormos (2020), who used objective linguistic

assessments, reported that most of the variance in comprehensibility was explained by fluency features, with pronunciation and grammatical accuracy playing a minimal role. In contrast, Saito et al. (2017), who employed subjective linguistic assessments, revealed that pronunciation and fluency together explained a large proportion of comprehensibility variance, while lexical and grammatical accuracy and sophistication played a role to a similar degree. Given these variations, the method of linguistic assessments may impact the linguistic correlates of comprehensibility.

Chapter 3 Method

3.1 The Purpose and Research Questions

The prior research consistently showed that comprehensibility is associated with various linguistic features—pronunciation, fluency, vocabulary, grammar, and discourse. Moreover, speech features like pronunciation and fluency greatly influenced comprehensibility compared to lexical, grammatical, and discourse-related features. However, some disparities arose when examining the effects of specific linguistic features on comprehensibility.

The current study aims to address this research gap by investigating whether linguistic features associated with comprehensibility vary depending on the subjectivity of linguistic assessments. To explore this, this study utilizes two distinct methods for linguistic assessments: subjective linguistic assessments employing 1000-point sliding scales, as developed by Saito et al. (2017), and objective linguistic assessments using widely accepted linguistic indexes obtained through acoustic analyses and automated evaluation systems known as the *Coh-Metrix* and *P_Lex* software.

The primary aim of this study is to identify linguistic features that influence comprehensibility, as consistently observed in both subjective and objective linguistic assessments. If this goal is accomplished, these linguistic features should be given

priority for potential inclusion in speaking curriculum and classroom instruction. In this thesis, I call subjectively assessed linguistic features "subjective linguistic features," and objectively assessed features "objective linguistic features."

The present study selected nine subjective linguistic features in the domains of pronunciation, fluency, grammar, and vocabulary. These features were drawn from Saito et al. (2017), which serves as the foundation of the current research. It is important to note that the prior study included two discourse-related features: cohesion and richness. However, this study chose not to incorporate these features because the previous research acknowledged the difficulty of assessing discourse features in short L2 speech samples, thus presenting a challenge for evaluation.

In terms of objective linguistic evaluations, 11 linguistic features were drawn from Trofimovich and Isaacs (2012), which correspond to the nine subjective linguistic features mentioned above.

The disparity in the number of linguistic variables between subjective and objective features arose from differences in their rating criteria. In subjective assessments, the evaluation of segmental errors encompasses the consideration of errors in individual sounds (i.e., vowels and consonants) and syllable structure (Saito et al., 2016). In contrast, in objective assessments, Trofimovich and Isaacs (2012) treated

these error types as separate two indexes. Similarly, for subjective assessments, the rating criteria for lexical richness involve an assessment of both lexical diversity and sophistication (Saito et al., 2016). On the other hand, for objective assessments, Trofimovich and Isaacs (2012), analyzed these features as distinct two indexes.

In alignment with these procedures, the current study treated segmental errors and syllable structure errors as two distinct indexes in the objective linguistic assessments, which correspond to a single index of segmental error in subjective assessments. Likewise, lexical diversity and sophistication were separately analyzed as two distinct indexes in the objective linguistic assessments in the current study, which align with a single index of lexical richness in subjective assessments. Consequently, the number of subjective linguistic features amounts to nine, and that of objective linguistic features totals 11 (i.e., nine plus two). The correspondence of the subjective and objective features is summarized in Table 1.

Table 1

Correspondence of Subjective and Objective Linguistic Features

Domain	Subjective features	Objective features
Pronunciation	1. Segmental error	1. Segmental error ratio
		2. Syllable structure error ratio
	2. Word stress error	3. Word stress error ratio
	3. Rhythm error	4. Vowel reduction error ratio
	4. Intonation error	5. Intonation error ratio
Fluency	5. Speech rate	6. Articulation rate
Lexis	6. Lexical appropriateness	7. Lexical error ratio
	7. Lexical richness	8. MTLD
		9. Lambda
Grammar	8. Grammatical error	10. Grammatical error ratio
	9. Grammatical complexity	11. Mean length of AS-unit

The study addresses the following research questions:

(1) In the domains of pronunciation, fluency, vocabulary, and grammar, which of the nine subjective linguistic features demonstrate significant relationships with comprehensibility ratings provided by L1 raters? Furthermore, which subjective linguistic features effectively differentiate L2 speakers at varying comprehensibility levels?

(2) In the domains of pronunciation, fluency, vocabulary, and grammar, which of the 11 objective linguistic features demonstrate significant relationships with comprehensibility ratings provided by L1 raters? Furthermore, which objective

linguistic features effectively differentiate L2 speakers at varying comprehensibility levels?

3.2 Speech Samples

The speech data set utilized in Mikami (2019) has been employed in the current study.

3.2.1 Speakers

Initially, 48 Japanese learners of English (16 men and 32 women) were involved. However, three participants were excluded due to either recording failure or insufficient speech data. Subsequently, the analysis was conducted with the remaining 45 participants (14 men and 31 women). The demographic characteristics and descriptive statistics for this cohort are presented in Table 2.

All the participants were undergraduate students majoring in diverse academic disciplines at Japanese universities. The mean age was 20.2 years, ranging from 18–22. Japanese was the L1 for all participants. Their English language proficiency spanned a broad spectrum, as indicated by the duration of their English language studies, with an average of 10.05 years and a range from 5.00–18.83 years. Furthermore, 28 of the participants had lived in English-speaking countries, with an average duration of 0.68 years, ranging from 0.08–6.75 years. Based on self-report data from 38 participants,

their Test of English for International Communication (TOEIC) scores varied from 450–960, with an average score of 745.92. Out of the seven participants who did not report their TOEIC scores, three participants disclosed their scores on the Test of English as a Foreign Language (TOEFL). The scores of these 41 participants in either TOEIC or TOEFL collectively indicated a broad spectrum of English language proficiency, spanning the A2 to C1 on the Common European Framework of Reference for Languages (ETS, 2008). The proficiency levels of the remaining four participants could not be identified through available data.

Table 2

Descriptive Statistics for Speakers' Background Information (N = 45)

Variable	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Age	20.2	1.04	20	18	22	−0.63	−0.39	0.15
Years of studying English	10.05	3.32	9	5	18.83	1.06	0.09	0.50
Years spent in English-speaking country[a]	0.68	1.24	0.33	0.08	6.75	4.17	17.68	0.23
TOEIC score[b]	745.92	130.64	760	450	960	−0.28	−0.96	21.19

Note. [a]$n = 28$, [b]$n = 38$

3.2.2 Material

To maintain consistency with prior research (Saito et al., 2017; Trofimovich & Isaacs, 2012), the same eight-frame picture narrative (Derwing et al., 2009) was employed as material for a speaking task. In this picture story, a man and a woman

unintentionally collide at a street corner while strolling along a bustling street in a large city. The collision causes them to drop their identical suitcases, which become mixed up. They realize that their respective luggage was switched after arriving at their residence.

3.2.3 Recording

The participants each engaged in an individual recording session in a sound-treated room. The author explained the study's purpose and procedural details to the participants before initiating the recording. The participants completed a consent form after understanding and agreeing to voluntary participation in the study. The author then proceeded with the recording session.

During the recording session, participants were introduced to the eight-frame narrative to familiarize themselves with its content and plan what they would say about the story. A one-minute preparation period was allocated for this task, consistent with prior studies (Isaacs & Trofimovich, 2012; Saito et al., 2017; Trofimovich & Isaacs, 2012). The participants were not allowed to take notes during this phase. Following the one-minute preparation interval, the participants were instructed to commence narrating the story without any time constraints.

All speech samples were recorded using a condenser microphone, the AKG C4000B, and a digital-analog converter, the Fireface UCX, at a sampling rate of 44.1kHz with 16-bit quantization. The recording software was Sound Forge Pro. These audio recordings were saved in WAV format onto an iMac computer.

Following the recording phase, the participants completed a questionnaire encompassing their personal details and language backgrounds. Finally, participants received a modest compensation for their participation. The entire session lasted approximately 30 minutes.

3.2.4 Preparation of Speech Samples and Transcriptions for Rating Tasks

The duration of speeches varied considerably, ranging from 24.4–246.3 seconds across speakers. Aligned with prior studies (Isaacs & Trofimovich, 2012; Saito et al., 2017; Trofimovich & Isaacs, 2012), the author extracted short excerpts from the original speech samples for subsequent comprehensibility assessments and linguistic analyses. Speech analysis software (Praat Version 6.0.16) (Boersma & Weenink, 2016) was used for the editing process.

Initially, disfluency markers found at the beginning of the original samples, such as filled pauses, silent pauses, and false starts, were removed. Subsequently, approximately 30 seconds were selected from the beginning of the edited samples,

guided by waveform and pitch contour analysis, ending at a natural pause and syntactic boundary. Finally, these edited excerpts underwent amplitude normalization to ensure uniform sample volume levels.

Following the editing phase, all auditory samples were exactly transcribed in standard orthography (i.e., unpruned transcriptions) by the author, enabling the evaluation of lexical and grammatical features. A second coder, a graduate student specializing in English Education, independently transcribed 18 randomly selected samples (40% of the total samples) to verify the author's transcription accuracy. Inter-coder agreement was assessed using Cohen's κ coefficient, yielding a substantial agreement level of $\kappa = .912$ ($p < .001$), indicating near-perfect consensus between the two transcribers (Landis & Koch, 1977). Furthermore, pruned transcriptions were created by removing disfluency markers found in the middle of unpruned transcriptions, such as filled pauses, verbatim repetitions, and self-corrections. In addition, these pruned transcriptions were corrected for English accuracy by a native English speaker, a graduate student specializing in English Education, to calculate the lexical error ratio and grammatical error ratio in the objective linguistic evaluations (see Section 3.4.2). Moreover, all unpruned textual data were transcribed into the International Phonetic Alphabet (IPA), per the conventions outlined by Wells (2008), to assess pronunciation

and fluency features. A phonetically trained coder, an English teacher at a private school who holds a master's degree in English Education with a specialization in English Phonetics, independently transcribed 18 files using the same procedural methodology as the orthographic transcriptions to verify the author's IPA transcription accuracy. Inter-coder agreement was assessed, resulting in substantial agreement with a Cohen's κ coefficient of .780 ($p < .001$), affirming substantial concordance between the two transcribers (Landis & Koch, 1977). Tables 3 and 4 summarize the descriptive statistics for speech duration in the complete and extracted speech samples and the word count for unpruned and pruned transcriptions.

Table 3

Descriptive Statistics for Speech Duration (Seconds) in Original and Edited Speech Samples

Variable	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Complete	92.86	46.06	90.70	24.40	246.30	0.87	1.04	6.87
Excerpted	30.84	4.31	31.60	20.60	37.30	−0.34	−0.86	0.64

Table 4

Descriptive Statistics for Word Count in Unpruned and Pruned Transcriptions

Variable	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Unpruned	42.20	13.38	41.00	17.00	80.00	0.47	0.21	1.99
Pruned	37.71	13.19	38.00	16.00	77.00	0.63	0.18	1.97

3.3 Comprehensibility Ratings

3.3.1 Raters

A cohort of 10 native speakers of North American English (four men and six women) participated as comprehensibility raters: four raters in Mikami (2019) (i.e., their data) were reused and six new raters were additionally recruited. Self-reported background information is encapsulated in Table 5.

All raters were enrolled as exchange students at a Japanese university, with an average age of 22.9. All raters self-identified English as their L1, except for one rater who designated Swedish as their L1. This rater was included in further analysis due to the results from an extra interview. Specifically, this rater relocated to the United States at six, receiving education in American elementary and college schools for 12 years. Notably, the rater predominantly employed English daily, except for interactions with the rater's parents. All raters demonstrated a reasonable degree of familiarity with the Japanese language, evidenced by their mean duration of their study of the Japanese language (4.06 years). Furthermore, the raters exhibited a high level of familiarity with Japanese-accented English, registering an average score of 4.8 on a 6-point scale where one indicated a low degree of familiarity and six indicated a high degree of familiarity. All raters reported no hearing problems.

Table 5

Descriptive Statistics for Comprehensibility Raters' Background (N = 10)

Variable	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Age	22.9	3.18	22	19	27	0.27	−1.78	1.00
Years of staying in Japan	0.85	0.99	0.33	0.08	3.17	1.19	0.13	0.31
Years of studying Japanese	4.06	2.87	3.58	0.50	10.42	0.83	−0.22	0.91
Familiarity with Japanese-accented English[a]	4.8	1.32	5.0	2.0	6.0	−0.73	−0.64	0.42

Note. [a]1 = Not at all familiar, 6 = Very familiar.

3.3.2 Procedure

All participants took part individually in the session for rating comprehensibility, conducted in a sound-treated studio environment. The author explained the objectives and procedures before initiating the rating session. The participants completed consent forms, signifying their understanding and voluntary participation in the current study. The practice session then commenced. During the practice session, the participants were initially presented with the eight-frame visual narrative used to elicit speech samples. The purpose was to familiarize them with the material they would later listen to. This familiarization step was crucial to prevent potential biases, wherein raters might inadvertently attribute lower comprehensibility scores to earlier-presented speech samples due to their unfamiliarity with the content.

Subsequently, the raters were provided with the definition of comprehensibility (see Section 2.1) and were instructed to assess comprehensibility based on the following guidelines:

> [H]ow much effort it takes to understand what someone is saying. If you can understand with ease, then a speaker is highly comprehensible. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility (Saito & Akiyama, 2017, p. 217).

Participants underwent a practice phase to ensure the consistency of the rating procedure. In this practice session, they listened to three speech samples once only. These practice samples were collected in a pilot study and not included in the main rating session. After a single exposure, participants evaluated the comprehensibility of these samples using a 9-point scale, where 1 = easy to understand and 9 = difficult to understand. The ratings were assigned by selecting a single number from the horizontally presented options on the screen (see Figure 1). The order of presentation of speech samples was randomized for each participant. The evaluation was conducted using a program made with Hot Soup Processor (HSP) programming software. Upon

completing the practice phase, participants were given a designated period to seek clarification or raise questions about the session.

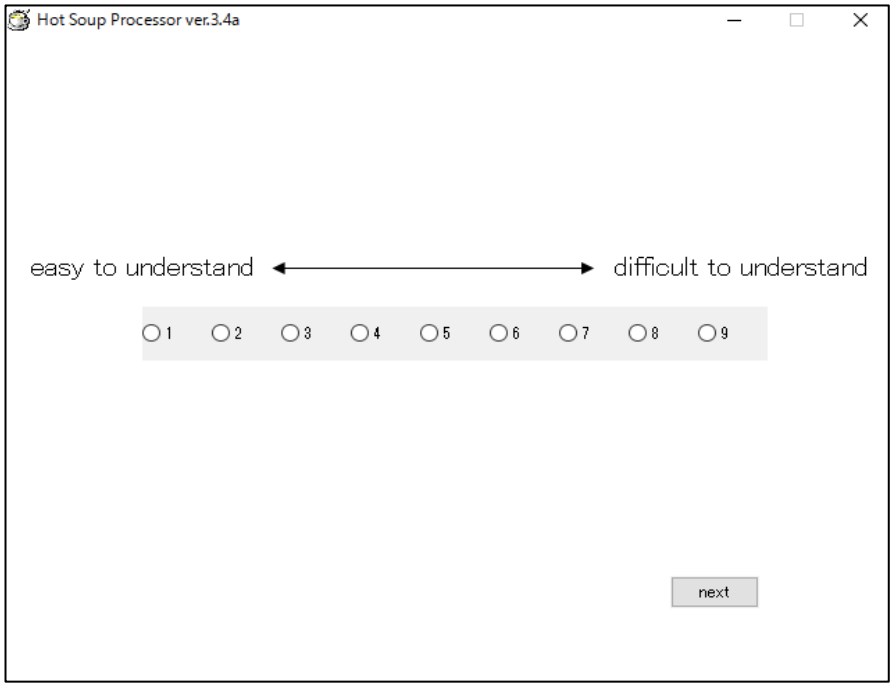
Subsequently, participants proceeded to the main rating session. This phase involved the evaluation of the comprehensibility of 45 speech samples. The procedure mirrored the practice session, allowing participants to evaluate the samples at their own pace.

Upon completing the rating session, participants completed a questionnaire regarding their background information. Additionally, they were given modest compensation for their participation. The session lasted approximately one hour.

The Cronbach's α value for comprehensibility attained a high value of .94, 95% CI [.92, .96], exceeding the acceptable threshold of .70–.80 (Larson-Hall, 2010). Accordingly, the scores provided by 10 raters were averaged to derive a single score for each speaker following prior investigations (e.g., Isaacs & Trofimovich, 2012; Saito et al., 2017). These averaged scores were utilized for further analysis.

Figure 1

A sample of On-screen Labels for Comprehensibility Assessments



3.4 Linguistic Analysis

3.4.1 Subjective Assessment

3.4.1.1 Linguistic Variables.

This study selected nine subjective linguistic features spanning the domains of pronunciation, fluency, grammar, and vocabulary to explore the linguistic features associated with comprehensibility. These features were drawn from prior investigations to ensure comparability (Crowther, Trofimovich, Isaacs, & Saito, 2015; Crowther, Trofimovich, Saito, & Isaacs, 2015; Saito et al., 2016; 2017; Saito & Shintani, 2015). As mentioned above, two discourse-related features (richness and cohesion) were not

included in the current study due to difficulty in observing these features in short speech samples (see Section 3.1).

Raters were presented with the following explanations (Crowther et al., 2015, pp. 86–87; Saito et al., 2016, pp. 235–236) to guide their assessment of each linguistic aspect.

1. ***Segmental Errors*** are errors in individual sounds. For example, if someone says "road" or "rain," but the listener hears an "l" sound instead of an "r" sound, that is a speaker's consonant error. If someone says "fan" or "boat," but the listener hears "fun" or "bought," that is a vowel error. Listeners may also hear sounds missing from words or extra sounds added to words. These are also consonant and vowel errors (Saito et al., 2016, p. 235).
2. ***Word Stress*** occurs when an English word has more than one syllable. One of the syllables will be slightly louder and longer than the others. For example, if the speaker says "computer," the listener may notice that the second syllable has more stress (comPUter). If the listener hears stress on the wrong syllable or equal stress on all of the syllables in a word, then these are word stress errors (Saito et al., 2016, p. 235).

3. ***Intonation*** can be thought of as the melody of English. It is the natural pitch changes that occur when we speak. For example, a listener may notice that the speaker's pitch goes up at the end of a question when they ask a question with a yes/no answer. If someone sounds "flat" when they speak, it is likely because their intonation is not following English intonation patterns (Saito et al., 2016, p. 235).
4. ***Rhythm*** refers to the difference in stress (emphasis) between content and function (grammatical) words. For instance, in the sentence My SISTER WORKS in an OFFICE, the words *sister*, *works*, and *office* are content words and therefore usually stressed more than the words *my*, *in*, and *an*, which are grammatical words featuring reduced vowels (Crowther et al., 2015, pp. 86–87).
5. ***Speech Rate*** is how quickly or slowly someone speaks. Speaking very quickly can make speech difficult to follow, but speaking too slowly can have the same effect. An appropriate speech rate should sound natural and be comfortable to listen to (Saito et al., 2016, p. 235).
6. ***Lexical Appropriateness*** refers to the appropriateness of the words used by the speaker. "If the speaker uses incorrect or inappropriate words, including words from the speaker's native language, lexical accuracy is low. Conversely, lexical accuracy is high if the speaker has all the lexical items required to accomplish the speaking

task and does so using frequently-used and/or precise lexical expressions" (Saito et al., 2016, p. 236)

7. ***Lexical Richness*** also refers to the sophistication and suitability for the demands of the speaking task. "The speech lacks lexical richness if the speaker uses a few simple, unnuanced words. However, if the speaker's language is characterized by varied and sophisticated uses of English vocabulary, the speech is lexically rich" (Saito et al., 2016, p. 236).
8. ***Grammatical Accuracy*** refers to "the number of grammar errors the speaker makes, including errors in word order and morphological ending" (Saito et al., 2016, p. 236).
9. ***Grammatical Complexity*** concerns the complexity and sophistication of the speaker's grammar. "Grammatical complexity is low if the speaker uses basic, simple, or fragmented structures or sentences. Grammatical complexity is high if the speaker uses elaborate, sophisticated grammar structures" (Saito et al., 2016, p. 236).

3.4.1.2 Raters

Five American raters (four men and one woman) participated in the present study to evaluate the above nine linguistic features: four raters in Mikami (2019) (i.e., their

data) were reused and one new rater was additionally recruited. The self-reported background information is available in Table 6.

All raters were L1 speakers of North American English, with an average age of 53.4. All raters had extensive experience in teaching English, averaging 24.4 years, within university or private school settings. Furthermore, four raters had studied linguistics, and three had master's degrees in English education (TEFL/TESOL). The selection of experienced native speakers was favored over less experienced counterparts, following the rationale that experienced speakers possess a heightened understanding of the concepts of the linguistic features under investigation, resulting in enhanced rating consistency among raters (Saito et al., 2017). Their mean duration of residence in Japan was 22.6 years and the average period of their Japanese language study was 12.2 years. The raters exhibited a high degree of familiarity with Japanese-accented English, as indicated by their consistent reporting of six on a 6-point scale, where 1 = not at all familiar, 6 = very familiar. All raters reported having no hearing problems.

Table 6

Descriptive Statistics for Linguistic Raters' Background (N = 5)

Variable	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Age	53.4	18.12	62	25	69	−0.58	−1.63	8.10
Years of studying Japanese	12.2	7.95	12	2.0	20	−0.10	−2.05	3.56
Years of staying in Japan	22.6	14.06	22	2.0	38	−0.29	−1.73	6.29
Years of teaching English	24.4	14.60	23	2.0	38	−0.43	−1.63	6.53
Familiarity with Japanese-accented English[a]	6.0	0.0	6.0	6.0	6.0	—	—	0.0

Note. [a]1 = Not at all familiar, 6 = Very familiar.

3.4.1.3 Procedure.

Five raters participated individually to analyze nine linguistic features. The author employed 1000-point free-moving sliders (Saito et al., 2017) to assess these linguistic features. The raters performed linguistic analysis by using a computer program developed with HSP programming software. The author constructed this rating program. The rating sessions were conducted in quiet offices or sound-treated studios over two days to prevent fatigue.

On the first day, raters evaluated five pronunciation and fluency features—segmental error, word stress error, rhythm error, intonation error, and speech rate (i.e., audio-based features). The author explained the study's purpose and procedural details to the raters before commencing the rating session. The raters completed a consent form

to signify their voluntary participation, understanding, and consent. They then engaged in a practice session.

During the practice session, the raters were initially given explanations for rating each linguistic feature (see Section 3.4.1.1) and the eight-frame picture used for a speech elicitation task to familiarize them with the narrative they would later listen to. Subsequently, they listened to and evaluated three practice speech samples for the five audio-based linguistic features. The practice samples were collected in a pilot study and were not included in the main rating session. In accordance with the guidance provided in Saito et al. (2017), raters were informed that the speech samples covered a broad spectrum of English language proficiency. They were also advised that even a minor adjustment of the slider could indicate a considerable variation in the rating. Consequently, they were encouraged to utilize the entire range of each rating scale and evaluate the five audio-based linguistic features while referring the rating guidelines (see Section 3.4.1.1).

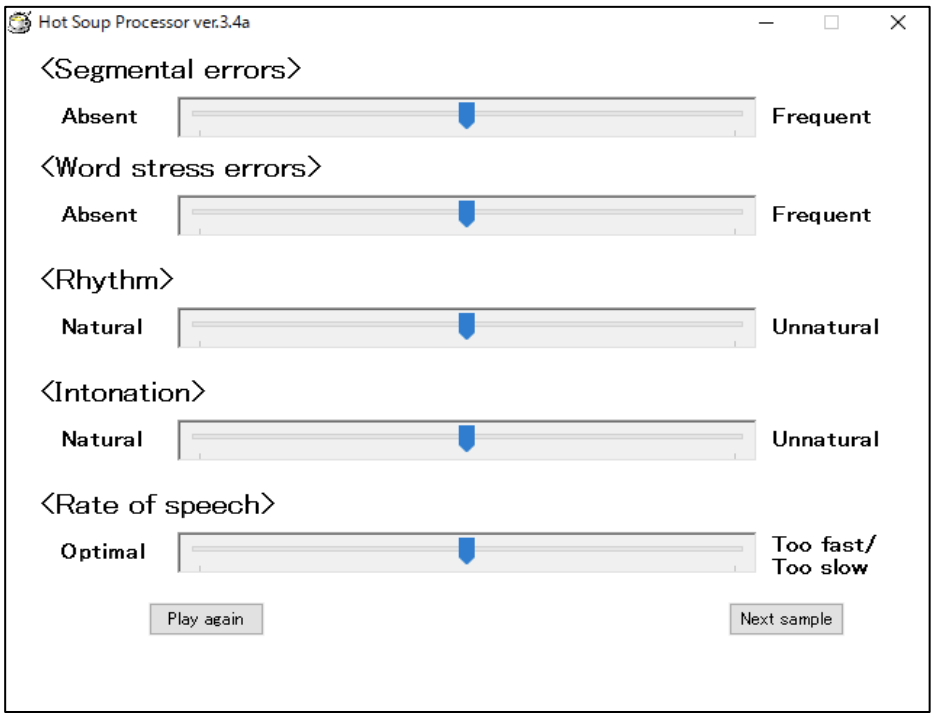
After the practice session, the raters could seek clarification or ask questions concerning the study and procedural details. Once satisfied with their understanding, they proceeded to the main session, assessing 45 speech samples for the five audio-based linguistic features. The order of presentation was randomized for each rater.

The following paragraphs provide detailed descriptions of the rating procedures. The raters initiated a speech sample by pressing the “Start” button at the bottom of the computer screen, the first sample automatically began playing, and five horizontal sliders appeared on the screen simultaneously (Figure 2). These sliders displayed the names of the five rating features (e.g., “segmental errors”) and brief descriptions of each rating category on the left and right sides of the sliders (e.g., “Absent” on the left side and “Frequent” on the right side for “segmental errors.”) The brief descriptions on the leftmost end indicated positive labels, while the rightmost end contained negative labels for all the features. The raters adjusted a blue tab, initially positioned in the middle of the slider, using a mouse while simultaneously listening to the speech samples to assess the five audio-based linguistic features. The raters could modify their ratings as often as needed during and after listening. Additionally, they could replay each speech sample as many times as desired by clicking the “Play again” button at the bottom left of the screen until they reached a final decision. Once they pressed the “Next sample” button at the bottom right of the screen, assessment values for all rating features were automatically recorded in the computer. For instance, if the tab was at the leftmost end, a rating value of 1 was recorded; if it was at the rightmost end, a rating value of 1000 was recorded. Then, the next speech sample automatically began playing, and the five

sliders were reset, and the tabs returned to the middle of the sliders. This process took approximately two hours on the first day.

Figure 2

A Sample of On-screen Labels for Linguistic Assessments



On the second day, the raters evaluated the remaining four lexical and grammatical features—lexical appropriateness, lexical richness, grammatical accuracy, and grammatical complexity (i.e., transcription-based features). Similar to the rating of pronunciation and fluency features on the first day, the raters were initially given explanations for rating each linguistic feature (see Section 3.4.1.1) and underwent a practice session. Consistent with the evaluations of pronunciation and fluency

mentioned earlier, raters were instructed that the speech samples encompassed a wide range of English language proficiency. They were also guided that even a small change in the slider position could signify a considerable difference in the rating. Hence, they were encouraged to use the entire range of each rating scale and evaluate the four transcription-based features while referring the rating guidelines (see Section 3.4.1.1). Furthermore, instead of listening to speech samples, the raters were informed that they would read transcriptions of these samples to assess the four lexical and grammatical features.

During the practice session, the raters assessed three practice samples for the four lexical and grammatical features using free-moving sliders. Unlike the pronunciation and fluency features, the raters judged the four lexical and grammatical features by reading transcriptions of the speech samples instead of listening to them. This was done to avoid any phonological errors interfering with the assessment of lexical and grammatical features, in line with previous studies (e.g., Saito et al., 2017). The detailed procedures mirrored those for pronunciation and fluency assessment. After the practice session, the raters could seek clarification or ask questions regarding the study and rating procedures. Once confident in their understanding, they proceeded to the main

session, where they read 45 transcriptions and evaluated them for the four lexical and grammatical features. The order of presentation was randomized for each rater.

Upon completing the assessments on the second day, the raters completed a questionnaire detailing their personal information, language background, and their level of understanding of the nine linguistic features on a 9-point scale, where 1 = I did not understand this concept at all, to 9 = I understand this concept well. The results indicated that all raters demonstrated a comprehensive understanding of all rating categories, with average scores ranging 8.2–9.0, as summarized in Appendix 1.

Furthermore, inter-rater reliability coefficients (Cronbach's α) for all nine features were computed using the *alpha* function within the *psych* package (Revelle, 2020) to ensure the consistency of the evaluations. The results are summarized in Table 7. The alpha values ranged from .80 to .90, surpassing the acceptable threshold of .70–.80 (Larson-Hall, 2010) for all features. Accordingly, the nine subjective linguistic scores provided by five raters were averaged to derive a single score for each speaker following prior investigations (e.g., Isaacs & Trofimovich, 2012; Saito et al., 2017). These averaged scores were utilized for further analysis. Finally, the raters received a small payment for their participation. The session on the second day took approximately 1.5 hours.

Table 7

Cronbach's α for the Nine Subjective Linguistic Features

Variable		95% CI	
	α	*LL*	*UL*
Segmental error	.81	.73	.90
Word Stress error	.83	.75	.91
Rhythm error	.83	.75	.91
Intonation error	.84	.77	.92
Speech rate	.87	.81	.93
Lexical appropriateness	.80	.72	.88
Lexical richness	.90	.86	.94
Grammatical accuracy	.84	.79	.90
Grammatical complexity	.90	.86	.94

Note. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.

3.4.2 Objective Assessment

3.4.2.1 Linguistic Variables.

This study chose 11 corresponding objective linguistic features—encompassing pronunciation, fluency, grammar, and vocabulary—to ensure comparability with the subjective linguistic features discussed earlier. Prior research guided this selection (Isaacs & Trofimovich, 2012; Saito et al., 2017; Trofimovich & Isaacs, 2012).

Pronunciation

1. ***Segmental Error Ratio*** refers to the proportion of phonemic substitutions (i.e., vowel and consonant errors) relative to the total number of segments articulated (Trofimovich & Isaacs, 2012). For instance, if the word *this* was pronounced as /dɪs/

instead of /ðɪs/, wherein a single phonemic error occurred, and the total number of segments articulated were three, then the resulting segmental error ratio would be 33.3%.

2. ***Syllable Structure Error Ratio*** is the proportion of errors attributed to vowel and consonant epenthesis (insertion) and elision (deletion) relative to the total number of syllables articulated (Trofimovich & Isaacs, 2012). For example, if a speaker pronounces *woman* as /ʊmən/ instead of /wʊmən/ due to the omission of the initial /w/, this would result in one syllable structure error, with a total of two syllables articulated. The resulting syllable structure error ratio would be 50%.
3. ***Word Stress Error Ratio*** is the proportion of polysyllabic words produced in which the primary stress is either misplaced or omitted, regarding the total number of polysyllabic words uttered (Trofimovich & Isaacs, 2012). To illustrate, suppose a speaker produces the sentence "*An accident has happened.*" as "*an ac-CI-dent has HAP-pened*." In this case, the word "*accident*" is stressed improperly, and the total number of polysyllabic words produced is two. Hence, the resulting word stress error ratio would be 50%.
4. ***Vowel Reduction Error Ratio*** is an accuracy index that captures the stress-timed characteristics of English rhythm (Deterding, 2001). This index is operationalized as

the ratio of correctly reduced syllables to obligatory vowel reduction contexts in polysyllabic and function words (Trofimovich & Isaacs, 2012). To illustrate, consider the sentence "*the WOman WALK to the CORner,*" which contains five obligatory contexts (denoted in lowercase letters). If a speaker mispronounces the word "*woman*" as *wo-MAN*, producing a full vowel, the resulting vowel reduction error ratio would be 20%.

5. ***Intonation Error Ratio*** refers to the number of erroneous pitch patterns at the end of phrases, specifically at syntactic boundaries, divided by the total number of expected pitch pattern instances (Trofimovich & Isaacs, 2012). The expected accurate pitch patterns are sourced from Sugimori et al. (1997). To illustrate, consider the sentence, "*After they went back home* [leveling pitch], *they found out that they had taken the wrong suitcase* [falling pitch]." In this example, there are two expected pitch patterns. If the speaker concluded the *after*-clause with a rising pitch, creating an incorrect pitch pattern, the resulting intonation error ratio would be 50%.

Fluency

Fluency is a multifaceted construct characterized by three distinct components: speed, breakdown, and repair (Skehan, 2003; Tavakoli & Skehan, 2005). Speed fluency

refers to the velocity at which a speaker delivers their speech. Breakdown fluency concerns the speaker's pausing patterns—such as the frequency and duration of pauses. Repair fluency encompasses the speaker's ability to correct or repeat words or phrases to improve their speech production. This study only employed speed fluency, as previous research had not verified the validity of using a 1000-point sliding scale to assess breakdown and repair fluency (Saito et al., 2017).

6. ***Articulation Rate*** is the number of pruned syllables uttered per second (Suzuki & Kormos, 2020). This rate is obtained by dividing the number of pruned syllables by the entire phonation time of the speech sample. Pruned syllables are those devoid of disfluencies, such as (un)filled pauses, repetitions, self-corrections, and false starts. Notably, the phonation time used in the calculation must exclude all disfluencies present in the speech sample.

Grammar

7. ***Grammatical Error Ratio*** refers to the proportion of words containing one or more morphosyntactic errors in relation to the total word count (Isaacs & Trofimovich, 2012). Morphosyntactic errors may manifest in various forms, such as inaccurate verb tense, morphology, or syntax. For instance, the following sentence, "*A man and woman were walking from different direction and bump into each other*," highlights

two morphosyntactic errors—a failure to observe the plural agreement in "*direction*," and the absence of the past-tense suffix "*ed*" in "*bump*." This passage contains 14 words and two grammatical errors. Thus, the resulting grammatical error ratio would be 14.2%.

8. ***Mean Length of AS-units*** is a widely used index for assessing grammatical complexity in L2 speech research (e.g., De Clercq & Housen, 2017; Tavakoli, 2018; Yu & Lowie, 2020). This index is computed by dividing the total number of words by the total number of AS-units, with higher values indicating greater grammatical complexity. An AS-unit is "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster et al., 2000, p. 365). For instance, consider the following utterance: "[*One salary man walked to the corner with his bag*.] [*And the woman also walked to the corner with the bag which is a same bag of the man*.]" This passage consists of 29 words and two AS-units (enclosed within square brackets). Consequently, the mean length of the AS-units would be 14.5.

Lexis

9. ***Lexical Error Ratio*** is the total number of incorrectly used lexical expressions divided by the total number of words produced in the text. Examples of inaccurate

lexical expressions include imprecise vocabulary choices, such as substituting *attack* for *bump into*, and L1 intrusions, which occur when a non-native speaker mistakenly integrates a word or phrase from their native language into the target language. An instance of this would be using *biru* instead of *building*.

10. ***The Measure of Textual Lexical Diversity (MTLD)*** is an index of lexical diversity computed by analyzing the type-token ratio (TTR) of a text. This index represents the average length of consecutive word strings in a text that maintains a constant TTR value (McCarthy & Jarvis, 2010). A higher MTLD value indicates greater lexical diversity in the text. While other commonly used indexes of lexical diversity, such as TTR and HD-D (McCarthy & Jarvis, 2007), are prone to produce smaller values with longer texts, MTLD is more robust to such sensitivity, particularly with relatively short texts (Koizumi & In'nami, 2012; McCarthy & Jarvis, 2010). See McCarthy and Jarvis (2010) for a more comprehensive calculation formula and methodological aspects. The figure presented in this study was computed using the *Coh-Metrix* software (McNamara et al., 2014).
11. ***Lambda*** is an index that measures the degree of lexical sophistication employed by L2 learners. The index mathematically transforms the frequency count of unusual or advanced words a learner uses in their production using the Poisson distribution.

Typically, the range of lambda values falls between 0 and approximately 4.5, and a higher figure indicates an increased level of lexical sophistication. In contrast to other indexes, such as Nation's Lexical Frequency Profile scores (Laufer & Nation, 1995), one of the key advantages of lambda scores is their limited sensitivity to text length. In addition, lambda scores exhibit reasonable stability with short texts (see Meara & Bell [2001] and Meara & Miralpeix [2017] for a more comprehensive calculation formula and methodological aspects). To calculate the figure, P_Lex Version 2.31 was used (Meara, 2018).

3.4.2.2 Raters.

The author conducted objective linguistic analyses of all 11 linguistic variables. The author has had substantial training in phonetics with skills of transcribing L2 speeches into the International Phonetic Alphabet (IPA) and is also teaching pronunciation courses at the university level. A second coder, a graduate student with substantial English teaching experience at high school and university levels, also participated in the study. Both are well-trained in the analysis of L2 speech.

3.4.2.3 Procedure.

The author analyzed 45 speech samples, examining 11 linguistic variables. Segmental error ratio, syllable structure error ratio, word stress error ratio, intonation

error ratio, vowel reduction error ratio, and articulation rate were quantified using the Praat software. Specifically, segmental error, syllable structure, vowel reduction, and articulation rate were evaluated by examining the previously mentioned IPA transcriptions (see Section 3.2.4). Word stress and intonation were assessed through visual and auditory analysis of pitch contour and waveform. Lexical error ratio and grammatical error ratio were computed by visually comparing learners' transcriptions with the corresponding transcriptions corrected by a native English graduate student majoring in English Education. The Mean length of AS-units was determined by manually counting AS-units and automatically calculating word counts using the RStudio software. Lastly, MTLD and lambda were automatically computed using the *Coh-Metrix* and *P_Lex* software, respectively.

The second coder then independently assessed 18 randomly selected speech samples (40% of the whole samples) for the same variables to verify the accuracy of the author's analyses. As shown in Table 8, Cronbach's alpha coefficients for all variables ranged from .97 to 1.00, indicating strong agreement between the coders. The exception was the word stress error ratio, which demonstrated lower consistency (.78) when compared to the other features. To address this inconsistency, the two coders engaged in discussions to establish a consistent rating criterion for this feature and conducted

recoding until an agreement was reached. Consequently, the rating values assigned by the author were used for further analyses.

Table 8

Cronbach's α for the 11 Objective Linguistic Features

Variable	α	95% CI	
		LL	*UL*
Segmental error ratio	.99	.98	1.00
Syllable structure error ratio	.99	.98	1.00
Word stress error ratio	.78	.65	.90
Vowel reduction error ratio	.99	.98	1.00
Intonation error ratio	.97	.94	1.00
Articulation rate	1.00	1.00	1.00
Lexical error ratio	.99	.97	1.00
Grammatical error ratio	1.00	1.00	1.00
Mean length of AS-unit	1.00	1.00	1.00

Note. CI = confidence interval; *LL* = lower limit; *UL* = upper limit.

3.5 Statistical Analysis

All data were statistically analyzed using RStudio Version 1.3.1093 (R Core Team, 2020). The specific packages and functions used for each analysis are detailed in their respective sections. The analyses involved five primary components: descriptive statistics, correlation analyses, principal component analyses, hierarchical multiple

regression analyses, and multivariate analyses of variance. Detailed procedures for each analysis are provided in the following sub-sections.

3.5.1 Descriptive Statistics

Descriptive statistics were calculated for comprehensibility ratings, the nine subjective, and 11 objective linguistic features. Additionally, their data distributions were examined by reviewing histograms and QQ plots and conducting Shapiro-Wilk's normality tests.

3.5.2 Correlation Analysis

Correlation analyses explored the associations between comprehensibility and specific linguistic features. As several subjective and objective linguistic features exhibited non-normal distributions (see Section 4.1), Spearman's rank order correlation coefficients were calculated between all variables. The *pairs.panels* function in the *psych* package version 2.0.12 (Revelle, 2020) was used to compute these correlations. The significant levels were set at $\alpha = .05$. Plonsky and Oswald's (2014) standards for small (.25), medium (.40), and large (.60) effect sizes were employed to interpret the magnitude of correlation coefficients.

3.5.3 Composite Scores of Linguistic Features

As indicated in Table 16 (see Section 4.2 below), some linguistic features exhibited high correlations ($r > .90$, Plonsky & Ghanbar, 2018), which can cause multicollinearity issues in subsequent multiple regression analyses. Furthermore, the ratio of the sample size ($N = 45$) to linguistic variables (i.e., nine and 11) fell below the recommended guideline (the ratio should not be less than 5:1, Plonsky & Ghanbar, 2018). To address these concerns, the present study employed an exploratory Principal Component Analysis (PCA) to reduce the number of linguistic variables and simplify the interpretation, following previous studies in this field (Crowther, Trofimovich, Isaacs, & Saito, 2015; Crowther, Trofimovich, Saito, & Isaacs, 2015; Saito et al., 2016; 2017).

Most prior research has identified two underlying factors through the PCA, namely pronunciation and lexicogrammar, from subjective linguistic features, encompassing pronunciation, fluency, lexis, and grammar (Crowther, Trofimovich, Isaacs & Saito, 2015; Crowther, Trofimovich, Saito & Isaacs, 2015; Saito et al., 2016; 2017). Therefore, this study hypothesized extracting two factors of pronunciation and lexicogrammar from the nine subjective linguistic features under scrutiny. For consistency, the study also conducted a PCA on the 11 objective linguistic features.

While no previous studies have undertaken the analysis on objective linguistic features, the current study hypothesized obtaining a similar two-factor solution. The analytical process of PCA followed prior literature (Brown, 2006; Fabrigar et al., 1999; Tabachnick & Fidell, 2013) to ensure methodological rigor and consistency.

3.5.3.1 Tests of Assumptions

The statistical assumptions of factorability were assessed through three procedures: (1) examining the correlation matrix of the linguistic features, (2) performing Bartlett's (1954) test of sphericity, and (3) calculating Kaiser–Meyer–Olkin's (KMO) measure of sampling adequacy (Kaiser, 1970; 1974).

Regarding correlation matrix, correlation coefficients of at least .30 or higher are required to satisfy the assumption (Tabachnick & Fidell, 2013).

Secondly, Bartlett's test assesses the factorability of the correlation matrix. A significant result ($p < .05$) indicates that the correlation matrix has sufficient factorability to reveal statistically meaningful factor(s) and satisfy the assumption. The *cortest.bartlett* function in the *psych* package version 2.0.12 (Revelle, 2020) was used to perform this test.

Finally, the overall index of Kaiser–Meyer–Olkin's measure of sampling adequacy (MSA) examines the appropriateness of conducting subsequent PCA. An

index value exceeding .50 is deemed acceptable for verifying the sampling adequacy (Kaiser & Rice, 1974; Shirkey & Dziuban, 1976). The *KMO* function in the *psych* package version 2.0.12 (Revelle, 2020) was utilized to calculate this index.

3.5.3.1.1 Subjective Features

As for the factorability of the nine subjective linguistic features, the correlations among all pairs (see Table 16 in Section 4.2 below) surpassed the threshold of .30, indicating meaningful factorability. Bartlett's test of sphericity ($\chi^2(36) = 679.56$, $p < .001$) showed that these correlations were sufficiently substantial for PCA. Furthermore, the overall Kaiser–Meyer–Olkin's measure (KMO = .89) confirmed the sampling adequacy, exceeding the acceptable threshold of .50 (Kaiser & Rice, 1974; Shirkey & Dziuban, 1976). In sum, the statistical assumption of factorability was successfully verified for the subjective linguistic features.

3.5.3.1.2 Objective Features

Regarding the factorability of the 11 objective linguistic features, out of the 55 pairs of features, seven exhibited significant correlations above the acceptable threshold of .30 (Tabachnick & Fidell, 2013, see Table 16 in Section 4.2 below). Second, Bartlett's test of sphericity, $\chi^2(55) = 172.18$, $p < .001$, confirmed that these correlations were sufficiently substantial for the subsequent PCA. Finally, the overall Kaiser–

Meyer–Olkin's measure confirmed the sampling adequacy for the analysis, yielding a value of KMO = .53, surpassing the acceptable benchmark value of .50 (Kaiser & Rice, 1974; Shirkey & Dziuban, 1976). Above all, the statistical assumption of factorability was met for the objective linguistic features.

3.5.3.2 Main Analysis

Following the confirmation of statistical assumptions, initial solutions of PCA were computed to ascertain the suitable number of factors for extraction. Conforming to the Kaiser–Guttman rule (Guttman, 1954; Kaiser, 1960), factors with eigenvalues exceeding one were considered significant and thus retained for further analyses.

In cases where the initial run of PCA resulted in multiple factors, a rotation was employed to enhance their interpretability. As justified by Fabrigar et al. (1999), oblique rotation was initially preferred over orthogonal rotation. Only if the solutions with oblique rotation indicated negligible correlations between the extracted factors (i.e., $r < .30$, as suggested by Tabachnick & Fidell, 2013), then a subsequent PCA accompanied by orthogonal rotation was performed iteratively. Following Brown (2006), factor loadings equal to or greater than .40 were deemed salient, signifying meaningful relationships between the features and the respective factors. The analyses

were conducted using the *principal* function in the *psych* package version 2.0.12 (Revelle, 2020).

3.5.3.2.1 Subjective Features

A PCA was conducted on the nine subjective linguistic features. An initial analysis was undertaken using a 9-factor solution without rotation, and eigenvalues for each component were obtained to determine the appropriate number of components to retain. The scree plot and eigenvalues for each component are presented in Figure 3 (factor loadings are summarized in Appendix 2). Figure 3 demonstrated that the first two components exhibited eigenvalues greater than 1, collectively accounting for 92.5% of the total variance. In light of these findings, the two components were deemed suitable for retention in the subsequent analysis.

Consequently, a second PCA was performed with a 2-factor solution. Because the factor loadings were not interpretable, (all nine features loaded on Factor 1, and two simultaneously loaded on Factor 2, see Appendix 3), Promax rotation (oblique rotation) was applied to the extracted two factors to obtain a more interpretable solution.

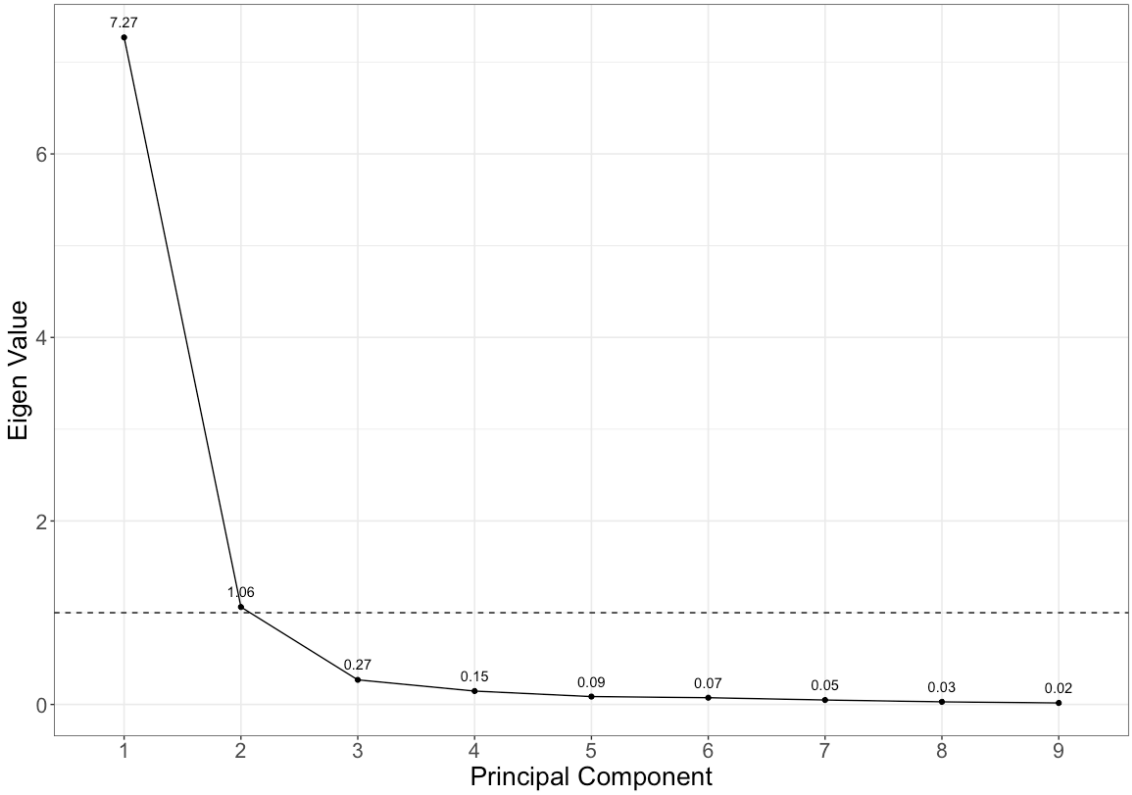
Table 9 displays the factor loadings with a two-factor solution followed by Promax rotation. The table revealed a distinct pattern, with five features of pronunciation and fluency loading on Factor 1 and the remaining four features of lexis

and grammar loading on Factor 2. This outcome aligns with a previous finding (Saito et al., 2017). Consistent with the prior study, Factor 1 was labeled as “pronunciation,” and Factor 2 was labeled as “lexicogrammar.” Notably, while Factor 1 is called “pronunciation,” it encompasses both pronunciation and fluency features. The correlation between pronunciation and lexicogrammar was $r = .71$, $p < .001$, 95% CI [.52, .83].

In addition to the PCA scores, composite scores of pronunciation and lexicogrammar were calculated using z-score-based scoring approach (Stanovich & West, 1989) to facilitate the comparison with objective linguistic features (see Section 3.5.3.3 below).

Figure 3

Scree plot for a Principal Component Analysis of the Nine Subjective Linguistic Features with a Two-factor Solution Followed by No Rotation



Note. The dashed line serves as a reference, indicating an eigenvalue of 1. The values displayed in the figure represent the eigenvalues corresponding to each factor.

Table 9

Results from a Principal Component Analysis of the Nine Subjective Linguistic Features with a Two-factor Solution Followed by Promax Rotation (Pattern Matrix)

Variable	Factor loading	
	1	2
Segmental error	**.96**	−.02
Word stress error	**1.00**	−.05
Rhythm error	**.98**	.00
Intonation error	**.96**	.02
Speech rate	**.71**	.28
Lexical appropriateness	.10	**.87**
Lexical richness	−.10	**1.04**
Grammatical accuracy	.21	**.79**
Grammatical complexity	−.04	**1.01**

Note. The factors were extracted with an oblique (Promax) rotation. Factor loadings above .40 are in bold.

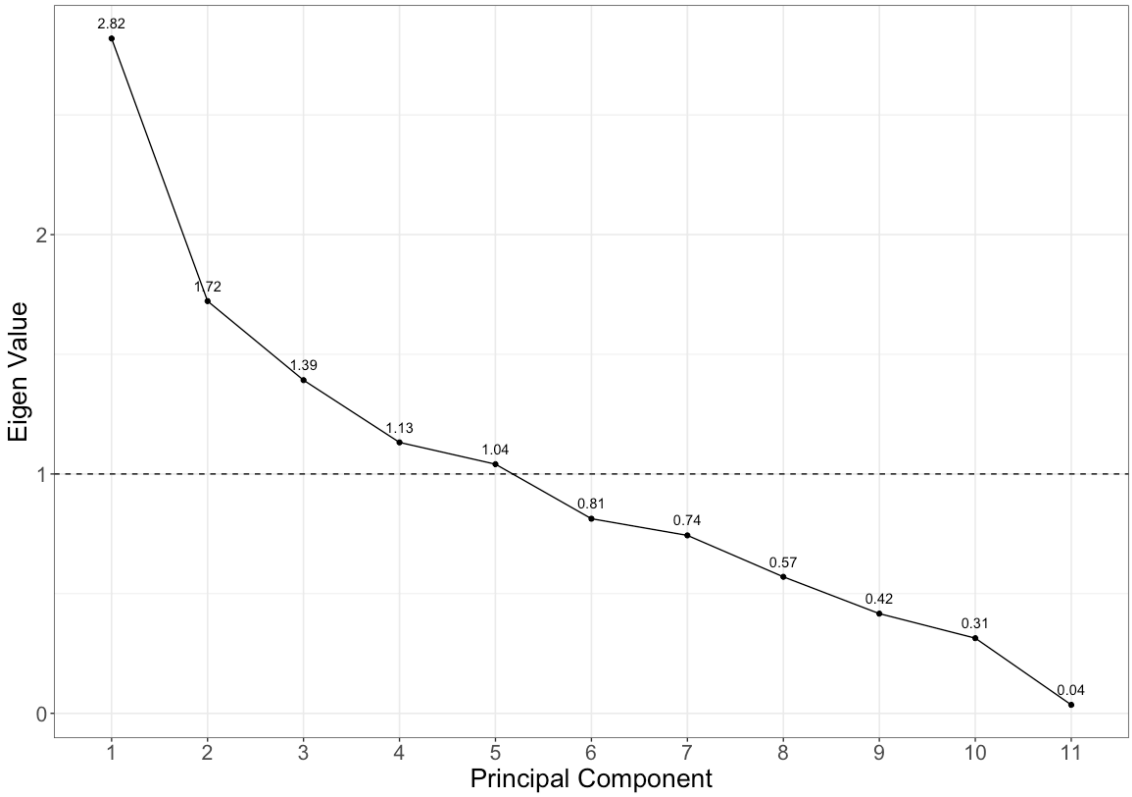
3.5.3.2.2 Objective Features

An initial solution with an 11-factor solution followed by no rotation was computed to determine the optimal number of principal components to extract. The scree plot and eigenvalues for each component are presented in Figure 4, along with factor loadings summarized in Appendix 4. Figure 4 indicated that the first five components possessed eigenvalues greater than 1, collectively accounting for 73.6% of the total variance. Based on these results, the five components were deemed suitable for retention in the subsequent analysis.

Table 10 presents the results of PCA with a five-factor solution followed by no rotation. Unfortunately, however, the solution was not interpretable. Accordingly, a series of PCA was conducted in an effort to obtain an interpretable result from the 11 objective features by altering the number of factors and factor rotation methods (these results are summarized in Appendixes 5 through 16). However, none of the analyses provided an interpretable result.

Figure 4

Scree Plot for a Principal Component Analysis of the 11 Objective Linguistic Features



Note. The dashed line serves as a reference, indicating an eigenvalue of 1. The values displayed in the figure represent the eigenvalues corresponding to each factor.

Table 10

Results from a Principal Component Analysis of the 11 Objective Linguistic Features With 5-factor Solution Followed by No Rotation

Variable	Factor loading				
	1	2	3	4	5
Phonemic substitution ratio	**.91**	.13	−.04	.22	−.12
Vowel reduction error ratio	**.90**	.23	.09	.27	−.09
Syllable structure error ratio	**.72**	**.40**	.15	.03	.03
Intonation error ratio	.14	**−.78**	.17	.32	−.03
Articulation rate	−.35	**.67**	−.21	.08	.31
Mean length of AS-unit	−.31	.12	**.69**	.17	−.12
Lambda	−.29	.29	**−.59**	.29	−.13
Grammatical error ratio	.23	−.31	**−.51**	.38	.39
MTLD	−.22	**.41**	**.42**	.39	.32
Word stress error ratio	**−.40**	−.18	.11	**.63**	.12
Lexical error ratio	.30	−.18	.15	−.32	**.79**

Note. The factors were extracted without rotation. Factor loadings above .40 are in bold.

3.5.3.3 Calculation of Z-score-based Composite Scores

Despite the uninterpretable results of PCA on objective linguistic features, reducing the number of linguistic features by creating composite variables was necessary to make comparisons with subjective linguistic features. Therefore, a mathematical transformation was undertaken using z-scores (Stanovich & West, 1989).

Two composite variables as were found in the subjective features (pronunciation and lexicogrammar) were created in the following way. First, z-scores of the 11 original objective features were obtained. The positive and negative signs in four features

(articulation rate, MTLD, lambda, and mean length of AS-unit) were reversed (i.e., multiplied by −1) because positive and negative values lead to opposite interpretations. For example, a higher z-score of segmental error ratio indicated more errors (more negative), while a higher z-score of articulation rate represented greater fluency (more positive). After this adjustment, six z-scores of pronunciation and fluency were averaged to obtain a composite score of pronunciation, and five z-scores of lexis and grammar were averaged to derive a composite score of lexicogrammar. Each composite score was labeled as either pronunciation or lexicogrammar following the results from the subjective linguistic features reported in the previous section. Unlike subjective features, the correlation between pronunciation and lexicogrammar was not significant (r = .28, p =.06, 95% CI [−.01, .53]). These composite scores were used as independent variables in subsequent regression analyses for objective linguistic features. To facilitate comparison with subjective linguistic features, the same procedure was undertaken to create z-score-based composite scores of the subjective linguistic features.

3.5.4 Hierarchical Multiple Regression Analysis

3.5.4.1 Tests of Assumptions

To assess the satisfaction of several statistical assumptions, regression diagnostic tests were performed (Field et al., 2012; Keith, 2019; Plonsky & Ghanbar, 2018;

Stevens, 2002; Tabachnick & Fidell, 2013). First, the normality of residuals was examined through the inspection of QQ-plots and histograms for standardized residuals, alongside the implementation of the Shapiro-Wilk normality test for standardized residuals. The residuals were expected to follow a normal distribution. Second, the linearity and homoscedasticity of residuals were assessed by plotting standardized residuals against the predicted values. The plot should exhibit a random arrangement of dots evenly dispersed around the zero point. Third, the presence of outliers was evaluated by examining Cook's distance—greater than one indicates the presence of outliers for each observation. Fourth, the independence of error terms was assessed using the Durbin-Watson test. A test statistic approaching two is considered more desirable, while a test statistic falling between one and three is deemed acceptable (Field et al., 2012). Finally, multicollinearity was scrutinized by calculating the variance inflation factor (VIF). An acceptable range for VIF is below 6–7 (Keith, 2019).

For subjective linguistic features, the results of the regression diagnostic tests confirmed the satisfaction of all statistical assumptions. The Shapiro-Wilk normality test was not significant ($W = 97$, $p = .35$), which indicated that the residuals followed a normal distribution. Visual inspections, including the QQ-plot and histogram in Appendix 17, also supported this finding. Regarding the linearity and homoscedasticity

of residuals, the scatter plot in Appendix 18 showed dots evenly dispersed around the zero point, suggesting that these assumptions were satisfied. Cook's distance was analyzed for each observation to examine the presence of outliers, as shown in Appendix 19. All observations exhibited Cook's distance values below 1, indicating the absence of outliers. For the independence of error terms, the Durbin-Watson test yielded a non-significant result with a test statistic of 1.75 (p = .43), confirming the independence of error terms. Lastly, multicollinearity was assessed through the VIF, yielding a VIF value of 2.01, which provided no evidence of multicollinearity.

For objective linguistic features, all statistical assumptions were satisfied. A Shapiro-Wilk normality test was administered, indicating that the residuals conformed to a normal distribution (W = .99, p = .99). The visual scrutiny of the QQ-plot and histogram concerning the residuals in Appendix 20 also confirmed that the distribution was normal. Concerning the linearity and homoscedasticity of residuals, the scatter plot depicted in Appendix 21 revealed a distribution of data points evenly dispersed around the zero point, satisfying these assumptions. The Cook's distance values for each observation were presented in Appendix 22, revealing that the values remained below the acceptable threshold of 1, indicating the absence of outliers. The Durbin-Watson test yielded a statistically significant result of 1.10 (p = .002). Notably, while a non-

significant result is expected to validate this assumption, Field et al. (2012) have stated that a test statistic between one and three remains acceptable. Accordingly, the present study determined that the independence of error terms assumption was met. Lastly, multicollinearity was assessed via the VIF, revealing a VIF value of 1.09, which indicated the absence of multicollinearity. In summary, a battery of regression diagnostic tests affirmed the appropriateness of the regression model.

3.5.4.2 Main Analysis

Two sets of hierarchical multiple regression analyses were conducted—for subjective and objective linguistic features—to assess the relative importance of pronunciation and lexicogrammar. The dependent variable was comprehensibility. These analyses used the *lm* function in the *stats* package version 4.0.3 (R Core Team, 2020).

Hierarchical multiple regression analysis is a statistical method wherein independent variables are sequentially added to the model one by one, and the resulting change in R^2 is deemed the degree of influence of the entered independent variable. However, caution is essential, as the order of variable entry can significantly impact the observed changes in R^2 (Keith, 2019). Frequently, variables entered earlier explain a

larger proportion of variance, potentially leading to erroneous conclusions that these early-entered variables exert greater influence on the dependent variable.

To address these concerns, the current study employed two models each for subjective and objective linguistic features. In the first model, pronunciation was entered as the initial independent variable, followed by lexicogrammar. In the second model, the order of variable entry was reversed. Initially, the relative significance of pronunciation and lexicogrammar was evaluated by comparing the R^2 values when these variables were introduced as the first predictors. This involved a direct comparison of pronunciation and lexicogrammar as initial predictors. Subsequently, the R^2 values linked to the inclusion of the second variables were scrutinized—the changes in the R^2 values attributed to the addition of pronunciation and lexicogrammar were compared as the second set of predictors. An *anova* function from the *stats* package version 4.0.3 (R Core Team, 2020) was employed to assess the statistical significance of this incremental change in explained variance.

3.5.5 Multivariate Analysis of Variance

Previous research has revealed that the effects of specific linguistic features on comprehensibility varies at different comprehensibility levels (Isaacs & Trofimovich, 2012; Saito et al., 2016). Multivariate analysis of variance (MANOVA) was utilized to

examine the group differences for each linguistic feature across varying comprehensibility levels.

The L2 speakers were divided into three sub-groups according to their comprehensibility scores. Specifically, L2 speakers with comprehensibility scores ranging from 1.0–3.6 were assigned to the *High* group (n = 11); those with scores ranging from 3.7–6.3 were assigned to the *Intermediate* group (n = 25); and those with scores ranging from 6.4–9.0 were assigned to the *Low* group (n = 9).

3.5.5.1 Tests of Assumptions

Before conducting MANOVA, statistical assumptions were examined: multivariate normality and the homogeneity of covariance matrices. The Shapiro-Wilk multivariate normality test was utilized to explore the normality assumption (the *mshapiro.test* function in the *mvnormtest* package version 0.1.9; Jarek, 2012), and Box's test was used to examine the homogeneity of covariance matrices (the *BoxM* function in the *MVTests* package version 2.1.1; Bulut, 2019) for each of the three groups. A p-value equal to or greater than the significance level ($p \geq .05$) was considered as indication that the assumptions of multivariate normality and homogeneity of covariance matrices were satisfied.

Regarding the statistical assumptions for the subjective linguistic features, the results from the Shapiro-Wilk multivariate normality tests revealed that the assumption was not met for all three groups: High ($W = .35$, $p < .001$), Intermediate ($W = .59$, $p < .001$), and Low ($W = .56$, $p < .001$). Box's test indicated that the covariance matrices for the linguistic features among the three comprehensibility levels were not homogeneous ($\chi^2(90) = 264.91$, $p < .001$).

As for the objective linguistic features, the Shapiro-Wilk multivariate normality test for Intermediate group revealed a deviation from the normality assumption ($W = .72$, $p < .001$). Test statistics could not be computed for High and Low groups due to a mathematical issue, wherein the inversion matrix approached 0. This arose because the number of dependent variables exceeded the number of participants in those groups. Hence, this study conservatively established that these groups did not meet the normality assumption. Box's test demonstrated that the covariance matrices for the linguistic features among the three comprehensibility levels were not equivalent ($\chi^2(132) = 745.34$, $p < .001$).

3.5.5.2 Main Analysis

Because statistical assumptions were not met, robust MANOVA based on ranking data (Choi & Marden, 1997; Wilcox, 2017) was performed for both subjective and

objective linguistic features, instead of parametric MANOVA. In this analysis, the comprehensibility level (Low, Intermediate, and High) was the independent variable, and the linguistic features were dependent variables. The *cmanova* function in the *WRS* package version 0.24 (Wilcox & Schönbrodt, 2014) was used to conduct the robust MANOVA. Kruskal-Wallis tests were followed if MANOVA results were significant. The *kruskal.test* function in the *stats* package version 4.0.3 (R Core Team, 2020) was used for this analysis. The significance levels were adjusted using the Bonferroni correction and set at α values = .005 and .004 for subjective and objective linguistic features, respectively. Additionally, if Kruskal-Wallis test results were significant, Mann-Whitney *U*-tests were performed to compare linguistic scores between the High and Intermediate groups, Intermediate and Low groups, and High and Low groups. The *wilcox_test* function in the *coin* package version 1.4.2 (Hothorn et al., 2006) was used for this analysis. The significance level was Bonferroni-corrected and set at α = .016 for both subjective and objective linguistic features.

Chapter 4 Results

4.1 Descriptive Statistics

Descriptive statistics and the results of Shapiro-Wilk normality tests are provided in Tables 11–15. The findings indicated that comprehensibility exhibited a normal distribution ($W = .96$, $p = .19$). However, among the nine subjective linguistic features, grammatical accuracy and grammatical complexity deviated from normality ($ps < .05$). For the 11 objective linguistic features, seven features (segmental error ratio, syllable structure error ratio, word stress error ratio, vowel reduction error ratio, intonation error ratio, MTLD, lambda) demonstrated non-normal distributions ($ps < .05$).

Table 11

Descriptive Statistics for Comprehensibility Ratings

Variable	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Comprehensibility	4.9	1.8	5.2	1.1	8.2	−0.34	−0.76	0.27

Note. 1 = easy to understand, 9 = difficult to understand.

Table 12

Descriptive Statistics for the Nine Subjective Linguistic Features

Variable	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Segmental error	518	176	527	75	876	−0.45	−0.05	26.27
Word stress error	443	151	449	76	795	−0.12	−0.18	22.52
Rhythm error	497	167	509	62	821	−0.53	−0.04	24.94
Intonation error	485	177	492	60	801	−0.39	−0.46	26.34
Speech rate	521	194	514	82	863	−0.24	−0.68	28.98
Lexical appropriateness	617	160	631	228	898	−0.59	−0.39	23.84
Lexical richness	640	199	635	208	962	−0.31	−0.70	29.67
Grammatical accuracy	687	166	741	271	957	−0.69	−0.42	24.70
Grammatical complexity	702	190	741	228	964	−0.63	−0.36	28.28

Note. 1 = target-like, 1000 = non-target-like

Table 13

Descriptive Statistics for the 11 Objective Linguistic Features

Variable	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Segmental error ratio	0.20	0.11	0.20	0.02	0.45	0.46	−0.92	0.02
Syllable structure error ratio	0.03	0.05	0.02	0	0.24	2.46	8.02	0.01
Word stress error ratio	0.22	0.21	0.17	0	0.8	0.99	0.21	0.03
Vowel reduction error ratio	0.23	0.22	0.18	0	0.72	0.81	−0.54	0.03
Intonation error ratio	0.49	0.32	0.50	0	1.00	0.01	−1.17	0.05
Articulation rate	3.08	0.43	3.02	2.42	4.24	0.42	−0.44	0.06
Lexical error ratio	0.07	0.05	0.06	0	0.19	0.55	−0.43	0.01
MTLD	32.13	13.10	29.85	14	71.68	0.82	0.14	1.95
Lambda	0.40	0.39	0.33	0	1.40	0.58	−0.73	0.06
Grammatical error ratio	0.13	0.07	0.13	0	0.26	0.06	−0.98	0.01
Mean length of AS-unit	7.70	1.58	7.83	5	11.67	0.35	−0.33	0.24

Table 14

Results from Shapiro-Wilk Normality Tests for the Nine Subjective Linguistic Features

Variable	*W*	*p*
Segmental error	.97	.46
Word stress error	.99	.99
Rhythm error	.97	.46
Intonation error	.97	.35
Speech rate	.98	.63
Lexical appropriateness	.95	.05
Lexical richness	.96	.28
Grammatical accuracy	.94	.02
Grammatical complexity	.94	.03

Table 15

Results from Shapiro-Wilk Normality Tests for the 11 Objective Linguistic Features

Variable	*W*	*p*
Segmental error ratio	.94	.02
Syllable structure error ratio	.71	< .001
Word stress error ratio	.87	< .001
Vowel reduction error ratio	.88	< .001
Intonation error ratio	.94	.02
Articulation rate	.96	.23
Lexical error ratio	.95	.06
MTLD	.93	.009
Lambda	.87	< .001
Grammatical error ratio	.96	.21
Mean length of AS-unit	.96	.27

4.2 Correlation Analysis

Table 16 displays a correlation matrix among the comprehensibility, nine subjective, and 11 objective linguistic features.

Concerning subjective linguistic features, all features exhibited significant correlations with comprehensibility, demonstrating large effect sizes. Consistent with prior findings (e.g., Saito et al., 2016; 2017), pronunciation and fluency features displayed slightly stronger correlations (ρ = .79–.85) than lexical and grammatical features (ρ = .70–.79).

Regarding objective linguistic features, five pronunciation and one grammatical features exhibited significant correlations with comprehensibility: segmental error ratio (.69), syllable structure error ratio (.63), word stress error ratio (−.31), intonation error ratio (.31), vowel reduction error ratio (.64), and grammatical error ratio (.38). The results confirmed the prior findings, indicating that speech-related features exhibited a stronger association with comprehensibility than lexical and grammatical features. However, it is noteworthy that these findings appear somewhat extreme, as none of the lexical and grammatical features, except for grammatical errors, displayed significant correlations with comprehensibility. Note that word stress errors displayed a contrasting

trend to other significant features, indicating that as word stress errors decreased, comprehensibility also decreased.

Lastly, concerning the correlations between subjective linguistic features and their objective counterparts (see Table 17), five subjective features exhibited significant associations: segmental error (.78, with segmental error ratio and .54 with syllable structure error ratio), rhythm error (.73, with vowel reduction error ratio), speech rate (−.33, with articulation rate), lexical richness (−.36, with MTLD and −.42 with lambda), and grammatical error (.36, with grammatical error ratio). It is essential to note that, despite articulation rate, MTLD, and lambda showing negative correlations, this is a natural occurrence. *Smaller* values for subjective assessments of these features signify more positive evaluations (an optimal speech rate and higher lexical sophistication), while *larger* values for their objective counterparts indicate more positive evaluations. In simpler terms, these negative correlations imply that highly evaluated subjective linguistic features also received high evaluations in corresponding objective ratings.

Table 16

Correlation Matrix for Comprehensibility, Nine Subjective, and 11 Objective Linguistic Features

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*
1. Comprehensibility	—																				
2. Segmental error	.79*	—																			
3. Word stress error	.83*	.93*	—																		
4. Rhythm error	.82*	.88*	.91*	—																	
5. Intonation error	.85*	.82*	.87*	.97*	—																
6. Speech rate	.80*	.76*	.80*	.90*	.92*	—															
7. Lexical appropriateness	.79*	.63*	.64*	.65*	.65*	.74*	—														
8. Lexical richness	.70*	.54*	.57*	.61*	.62*	.73*	.85*	—													
9. Grammatical accuracy	.77*	.72*	.70*	.73*	.71*	.78*	.85*	.84*	—												
10. Grammatical complexity	.72*	.55*	.59*	.62*	.64*	.73*	.85*	.97*	.85*	—											
11. Segmental error ratio	.69*	.78*	.74*	.77*	.74*	.62*	.50*	.47*	.63*	.49*	—										
12. Syllable structure error ratio	.63*	.54*	.59*	.70*	.73*	.70*	.46*	.42*	.49*	.41*	.67*	—									
13. Word stress error ratio	− .31*	− .19	− .17	− .26	− .30*	− .34*	− .32*	− .28	− .30*	− .26	− .33*	− .37*	—								
14. Intonation error ratio	.31*	.22	.22	.24	.24	.24	.36*	.24	.22	.26	.06	.07	.18	—							
15. Vowel reduction error ratio	.64*	.69*	.67*	.73*	.73*	.59*	.42*	.38*	.54*	.42*	.90*	.69*	− .22	.11	—						
16. Articulation rate	− .26	− .16	− .15	− .28	− .31*	− .33*	− .32*	− .40*	− .31*	− .43*	− .17	− .06	.15	− .51*	− .24	—					
17. Lexical error ratio	.24	.07	.14	.16	.20	.13	.20	.05	.10	.00	.12	.28	− .17	.07	.17	− .04	—				
18. MTLD	− .18	− .25	− .21	− .12	− .09	− .17	− .32*	− .36*	− .31*	− .35*	− .19	− .01	.32*	− .19	− .05	.29	.01	—			
19. Lambda	− .22	− .13	− .20	− .26	− .31*	− .34*	− .32*	− .42*	− .33*	− .40*	− .08	− .12	.02	− .15	− .15	.12	− .22	.13	—		
20. Grammatical error ratio	.38*	.33*	.30*	.31*	.28	.21	.25	.14	.36*	.13	.28	.05	.02	.25	.20	− .02	.13	− .07	.06	—	
21. Mean length of AS-unit	− .29	− .31*	− .31*	− .27	− .30*	− .21	− .15	− .18	− .22	− .22	− .20	− .13	.25	.03	− .14	.08	− .15	.26	− .19	− .25	—

Note. Subjective linguistic features are in bold. Objective linguistic features are in italics.

Table 17

Correlations Between Subjective and Objective Linguistic Features

Domain	Subjective features	Objective Features	ρ
Pronunciation	1. Segmental error	1. Segmental error ratio	.78*
		2. Syllable structure error ratio	.54*
	2. Word stress error	3. Word stress error ratio	−.17
	3. Rhythm error	4. Vowel reduction error ratio	.73*
	4. Intonation error	5. Intonation error ratio	.24
Fluency	5. Speech rate	6. Articulation rate	−.33*
Lexis	6. Lexical appropriateness	7. Lexical error ratio	.20
	7. Lexical richness	8. MTLD	−.36*
		9. Lambda	−.42*
Grammar	8. Grammatical error	10. Grammatical error ratio	.36*
	9. Grammatical complexity	11. Mean length of AS-unit	−.22

Note. *$p < .05$.

4.3 Composite Scores of Linguistic Features

4.3.1 Subjective Features

Tables 18 and 19 present the descriptive statistics for the PCA scores of the nine subjective linguistic scores and composite scores obtained through the z-score-based approach, respectively.

Table 18

Descriptive Statistics for the PCA Scores of the Nine Subjective Linguistic Features

Variable	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Pronunciation	0.00	1.00	0.08	−2.54	2.07	−0.41	−0.13	0.15
Lexicogrammar	0.00	1.00	0.17	−2.41	1.54	−0.61	−0.42	0.15

Table 19

Descriptive Statistics for the Composite Scores of the Nine Subjective Linguistic Features Extracted from Z-score-based Approach

Variable	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Pronunciation	0.00	0.96	0.01	−2.44	1.96	−0.42	−0.13	0.14
Lexicogrammar	0.00	0.96	0.21	−2.37	1.39	−0.67	−0.37	0.14

4.3.2 Objective Features

Table 20 provides the descriptive statistics for the composite scores of 11 objective linguistic features obtained through the z-score-based approach.

Table 20

Descriptive Statistics for the Composite Scores of the 11 Objective Linguistic Features Extracted from Z-score-based Approach

Variable	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Pronunciation	0.00	0.55	−0.01	−0.88	1.14	0.29	−0.80	0.08
Lexicogrammar	0.00	0.53	0.02	−0.95	1.16	0.00	−0.95	0.08

4.4 Hierarchical Multiple Regression Analysis

4.4.1 Subjective Features

Table 21 illustrates the findings from a hierarchical multiple regression analysis for subjective linguistic features wherein pronunciation served as the initial predictor, followed by lexicogrammar ($F(2, 42) = 100.1$, $p < .001$). In the first step, pronunciation alone accounted for 75.3% of the total variance in comprehensibility. Subsequently,

with the addition of lexicogrammar in the second step, the explained variance increased by 6.5%. The change in R^2 was statistically significant ($p < .001$).

In the subsequent analysis, the order of variable entry was reversed.

Table 22 presents the results of this analysis ($F(2, 42) = 100.1$, $p < .001$). In the first step, lexicogrammar accounted for 63.4% of the total variance in comprehensibility. Then, the variance in comprehensibility was explained with the addition of pronunciation in the second step, which added 18.4%. This incremental change was statistically significant ($p < .001$).

The variance explained by the initial variables in each model revealed that the influence of pronunciation on comprehensibility was 1.2 times greater than that of lexicogrammar (75.3% compared to 63.4%). Similarly, a parallel comparison for the second variables in each model demonstrated that pronunciation had a higher impact on comprehensibility—2.8 times greater than that of lexicogrammar (18.4% compared to 6.5%).

Table 21

Results from Hierarchical Multiple Regression Analysis for Subjective Linguistic Features, with Pronunciation as the Initial Predictor, Followed by Lexicogrammar Extracted via Principal Component Analysis

Variable	*B*	95% CI for *B*		*SE B*	β	R^2	ΔR^2
		LL	*UL*				
Step 1						.753***	.753***
Constant	4.89***	4.62	5.16	0.13	—		
Pronunciation	1.59***	1.31	1.87	0.13	.87***		
Step 2						.818***	.065***
Constant	4.89***	4.66	5.13	0.11	—		
Pronunciation	1.11***	0.77	1.45	0.16	.60***		
Lexicogrammar	0.67***	0.34	1.01	0.16	.37***		

Note. CI = confidence interval; *LL* = Lower Limit; *UL* = Upper Limit.
*$p < .05$. ** $p < .01$. ***$p < .001$.

Table 22

Results from Hierarchical Multiple Regression Analysis for Subjective Linguistic Features, with Lexicogrammar as the Initial Predictor, Followed by Pronunciation Extracted via Principal Component Analysis

Variable	*B*	95% CI for *B*		*SE B*	β	R^2	ΔR^2
		LL	*UL*				
Step 1						.634***	.634***
Constant	4.89***	4.56	5.22	0.16	—		
Lexicogrammar	1.46***	1.13	1.80	0.16	.80***		
Step 2						.818***	.184***
Constant	4.89***	4.66	5.13	0.11	—		
Lexicogrammar	0.67***	0.34	1.01	0.16	.37***		
Pronunciation	1.11***	0.77	1.45	0.16	.60***		

Note. CI = confidence interval; *LL* = Lower Limit; *UL* = Upper Limit.
*$p < .05$. ** $p < .01$. ***$p < .001$.

An additional set of multiple regression analyses was undertaken by using z-score-based composite scores as in the case of objective features (see Section 3.5.3.3). The statistical assumptions were met as well (Appendixes 23 through 25). The results are presented in Appendixes 26 and 27. The overall findings remained consistent (either using PCA scores or z-scores), demonstrating that pronunciation had a greater influence on comprehensibility than lexicogrammar, irrespective of the order of variable entry (Pronunciation exhibited an influence approximately 1.2–2.9 times greater than that of lexicogrammar).

4.4.2 Objective Features

Table 23 presents the findings from a hierarchical multiple regression analysis on objective linguistic features with pronunciation as the initial predictor, succeeded by lexicogrammar ($F(2, 42), = 25.21, p < .001$).

In the initial step, pronunciation exclusively accounted for 41.6% of the total variance in comprehensibility. In the subsequent step, lexicogrammar contributed 10.7% to the explanation. The incremental change in R^2 was statistically significant ($p = .002$).

In a subsequent analysis, the order of variable entry was reversed. The results of this analysis are presented in

Table 24 ($F(2, 42) = 25.12$, $p < .001$). In the primary step, lexicogrammar alone accounted for 24.7% of the total variance in comprehensibility. In the subsequent step, pronunciation added 27.6% to the explanation. This incremental change in the R^2 was statistically significant ($p < .001$).

The variance explained by the initial variables in each model revealed that the influence of pronunciation on comprehensibility was 1.7 times greater than that of lexicogrammar (41.6% divided by 24.7%). Similarly, the variance explained by the second variables in each model was compared, demonstrating that pronunciation's impact on comprehensibility was 2.6 times greater than lexicogrammar (27.6% divided by 10.7%).

To summarize, the impact of pronunciation on comprehensibility judgment was more substantial than lexicogrammar. Pronunciation exhibited an influence approximately 1.7–2.6 times greater than lexicogrammar.

Table 23

Results from Hierarchical Multiple Regression Analysis for Objective Linguistic Features, with Pronunciation as the Initial Predictor, Followed by Lexicogrammar Extracted via Z-score Transformation Approach by Stanovich and West (1989)

Variable	*B*	95% CI for *B*		*SE B*	β	R^2	ΔR^2
		LL	*UL*				
Step 1						.416***	.416***
Constant	4.89***	4.47	5.31	0.20	—		
Pronunciation	2.19***	1.41	2.97	0.38	.65***		
Step 2						.523***	.107**
Constant	4.89***	4.51	5.27	0.18	—		
Pronunciation	1.85***	1.11	2.58	0.36	.55***		
Lexicogrammar	1.22**	0.46	1.98	0.37	.35**		

Note. CI = confidence interval; *LL* = Lower Limit; *UL* = Upper Limit.

*$p < .05$. ** $p < .01$. ***$p < .001$.

Table 24

Results from Hierarchical Multiple Regression Analysis for Objective Linguistic Features, with Lexicogrammar as the Initial Predictor, Followed by Pronunciation Extracted via Z-score Transformation Approach by Stanovich and West (1989)

Variable	*B*	95% CI for *B*		*SE B*	β	R^2	ΔR^2
		LL	*UL*				
Step 1						.247***	.247***
Constant	4.89***	4.41	5.37	0.23	—		
Lexicogrammar	1.77***	0.86	2.69	0.45	.51***		
Step 2						.523***	.276***
Constant	4.89***	4.51	5.27	0.18	—		
Lexicogrammar	1.22**	0.46	1.98	0.37	.35**		
Pronunciation	1.85***	1.11	2.58	0.36	.55***		

Note. CI = confidence interval; *LL* = Lower Limit; *UL* = Upper Limit.

*$p < .05$. ** $p < .01$. ***$p < .001$.

4.5 Multivariate Analysis of Variance

4.5.1 Subjective Features

Table 25 presents descriptive statistics for the nine subjective linguistic features at the *High*, *Intermediate*, and *Low* comprehensibility levels. Due to space limitations, the Intermediate group has been labeled as "Mid" in the table (also in Table 29, which provides descriptive statistics for objective linguistic features at different comprehensibility levels). The descriptive statistics indicates that the mean scores for all linguistic features improved with increasing comprehensibility levels.

Subsequently, a robust MANOVA was performed, revealing a significant main effect of comprehensibility on the linguistic features ($H(18) = 46.76$, $p < .001$). Accordingly, nine sets of Kruskal-Wallis tests were conducted, demonstrating significant effects of comprehensibility on all the linguistic features, with large effect sizes (Plonsky & Oswald, 2014) observed for each feature: segmental error, $H(2) = 24.80$, $p < .001$, $r = .69$, 95% CI [.49, .82]; word stress error, $H(2) = 26.81$, $p < .001$, $r = .72$, 95% CI [.54, .83]; rhythm error, $H(2) = 25.99$, $p < .001$, $r = .70$, 95% CI [.52, .83]; intonation error, $H(2) = 27.62$, $p < .001$, $r = .73$, 95% CI [.55, .84]; speech rate, $H(2) = 24.13$, $p < .001$, $r = .68$, 95% CI [.48, .81]; lexical appropriateness, $H(2) = 22.60$, $p < .001$, $r = .65$, 95% CI [.44, .79]; lexical richness, $H(2) = 19.67$, $p < .001$, r

= .60, 95% CI [.37, .76]; grammatical accuracy, $H(2) = 23.11$, $p < .001$, $r = .66$, 95% CI [.45, .80]; grammatical complexity, $H(2) = 22.44$, $p < .001$, $r = .65$, 95% CI [.44, .79].

Finally, Mann-Whitney's *U* tests were carried out to compare the difference of linguistic scores across the comprehensibility levels. Tables 26 through 28 present test statistics (*Z*), results from significance tests, effect sizes (*r*), and their corresponding 95% confidence intervals. The results showed significant differences in all linguistic scores between all pairs. Figure 5 summarizes the effect sizes when comparing the high and intermediate and the intermediate and low groups.

Table 25

Descriptive Statistics for the Nine Subjective Linguistic Features at High, Mid, and Low Comprehensibility Levels

Variable	Level	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Segmental error	High	328	159	369	75	533	− 0.28	− 1.49	47.93
	Mid	528	108	527	322	708	− 0.06	− 1.23	21.56
	Low	722	87	716	593	876	0.24	− 1.16	28.92
Word stress error	High	276	114	275	76	463	− 0.12	− 1.08	34.52
	Mid	449	91	449	267	671	0.12	− 0.22	18.24
	Low	628	86	623	519	795	0.52	− 0.94	28.63
Rhythm error	High	300	137	333	62	455	− 0.37	− 1.39	41.44
	Mid	521	100	543	304	689	− 0.29	− 0.74	19.99
	Low	675	98	667	482	821	− 0.43	− 0.66	32.60
Intonation error	High	265	119	291	60	426	− 0.30	− 1.39	35.93
	Mid	518	114	516	292	801	0.10	− 0.24	22.81
	Low	664	96	661	459	792	− 0.67	− 0.14	31.95
Speech rate	High	292	132	321	82	484	− 0.07	− 1.49	39.92
	Mid	550	125	547	321	782	0.03	− 0.74	24.99
	Low	719	143	747	394	863	− 1.11	0.21	47.59
Lexical appropriateness	High	432	142	400	228	656	0.28	− 1.46	42.91
	Mid	644	110	628	370	898	− 0.19	0.34	22.01
	Low	768	65	756	659	852	− 0.13	− 1.49	21.66
Lexical richness	High	426	167	388	208	739	0.30	− 1.26	50.48
	Mid	670	148	642	398	952	0.06	− 0.77	29.55
	Low	817	129	799	559	962	− 0.58	− 0.84	42.98
Grammatical accuracy	High	493	150	462	271	746	0.27	− 1.38	45.12
	Mid	716	110	741	470	916	− 0.49	− 0.49	22.02
	Low	844	66	867	755	957	0.09	− 1.36	21.88
Grammatical complexity	High	482	171	458	228	778	0.17	− 1.26	51.66
	Mid	736	129	743	471	944	− 0.14	− 0.80	25.84
	Low	877	81	880	741	964	− 0.29	− 1.61	27.02

Note. 1 = target-like, 1000 = non-target-like

Table 26

Results from Mann-Whitney's U Tests for the Nine Subjective Linguistic Features of the High and the Intermediate Groups

Variable	*Z*	*p*	*r*	95% CI
Segmental error	−3.30	< .001	.55	.27, .74
Word stress error	−3.59	< .001	.60	.34, .77
Rhythm error	−3.97	< .001	.66	.42, .81
Intonation error	−4.17	< .001	.70	.48, .83
Speech rate	−4.05	< .001	.68	.45, .82
Lexical appropriateness	−3.45	< .001	.58	.30, .76
Lexical richness	−3.55	< .001	.59	.33, .77
Grammatical accuracy	−3.66	< .001	.61	.35, .78
Grammatical complexity	−3.59	< .001	.60	.34, .77

Note. α = .016

Table 27

Results from Mann-Whitney's U Tests for the Nine Subjective Linguistic Features of the Intermediate and the Low Groups

Variable	*Z*	*p*	*r*	95% CI
Segmental error	3.75	< .001	.64	.39, .81
Word stress error	3.88	< .001	.67	.42, .82
Rhythm error	3.26	< .001	.56	.27, .75
Intonation error	3.38	< .001	.58	.30, .77
Speech rate	2.91	.003	.50	.19, .72
Lexical appropriateness	3.08	.001	.53	.23, .74
Lexical richness	2.46	.012	.42	.10, .67
Grammatical accuracy	2.95	.002	.51	.20, .72
Grammatical complexity	2.95	.002	.51	.20, .72

Note. α = .016

Table 28

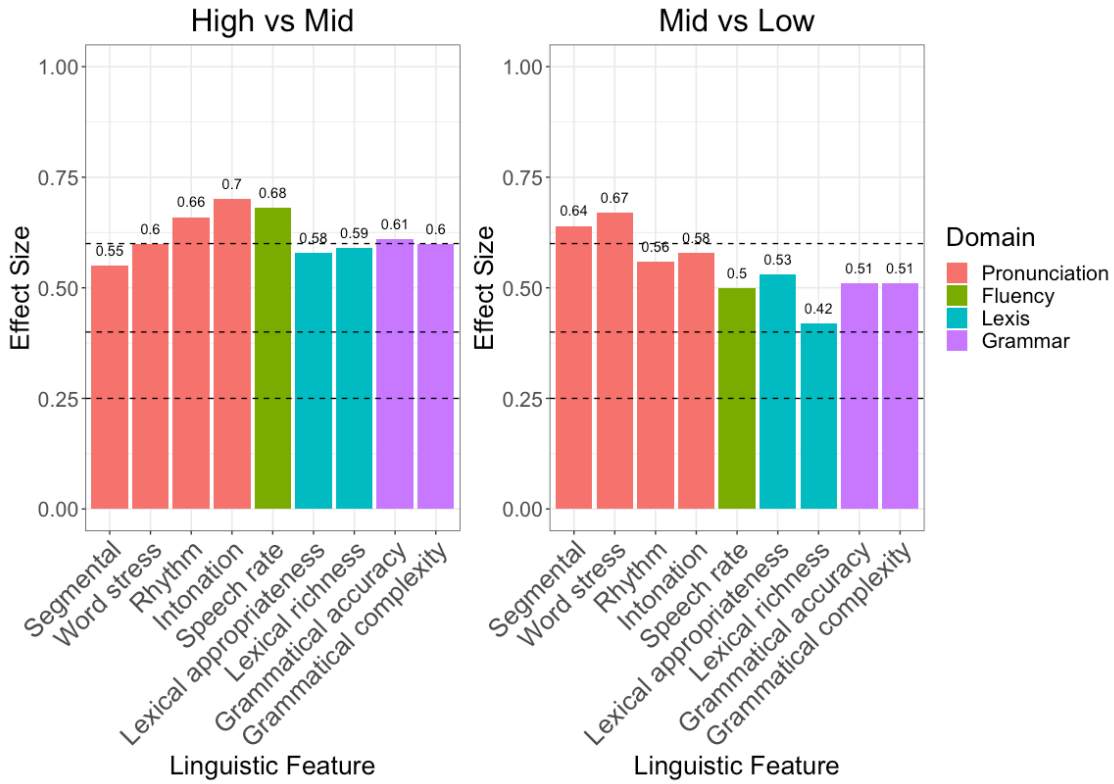
Results from Mann-Whitney's U Tests for the Nine Subjective Linguistic Features of the High and the Low Groups

Variable	*Z*	*p*	*r*	95% CI
Segmental error	−3.76	< .001	.84	.63, .94
Word stress error	−3.76	< .001	.84	.63, .94
Rhythm error	−3.76	< .001	.84	.64, .94
Intonation error	−3.76	< .001	.84	.63, .94
Speech rate	−3.53	< .001	.79	.53, .91
Lexical appropriateness	−3.76	< .001	.84	.63, .94
Lexical richness	−3.46	< .001	.77	.50, .91
Grammatical accuracy	−3.76	< .001	.84	.63, .94
Grammatical complexity	−3.68	< .001	.82	.60, .93

Note. $\alpha = .016$

Figure 5

Summary of Effect Sizes for Comparisons Between High and Intermediate Groups and Intermediate and Low Groups



Note. The dashed lines indicate reference points corresponding to effect sizes for small (.25), medium (.40), and large (.60) based on Plonsky and Oswald (2014).

First, in the effect sizes (*r*) comparing High and Intermediate comprehensibility levels, most of the linguistic features exceeded large effect sizes of .60, except for segmental error (.55), lexical appropriateness (.58), and lexical richness (.59), which also approached large effect sizes. These findings indicated that pronunciation and fluency features distinguished between High and Intermediate levels to a greater extent than did lexis and grammar features.

An interesting observation was found when comparing effect sizes among five pronunciation and fluency features. In these domains, the relatively larger units (rhythm and intonation accuracy, and speech rate) demonstrated larger difference between the levels when contrasted with the relatively smaller pronunciation units (segmental and word stress accuracy). Within the realms of lexis and grammar, all four features held equal power in distinguishing between High and Intermediate levels of comprehensibility.

Next, in the effect sizes comparing Intermediate and Low comprehensibility levels, four aspects of pronunciation exhibited a stronger discriminatory power (.56–.67) than lexical and grammatical features (.42–.53). These findings align with the outcomes of comparing High and Intermediate groups.

Notably, among the four pronunciation features, relatively smaller components (segmental and word stress accuracy) separated Intermediate from Low levels to a greater extent (.64–.67) than the relatively larger units (rhythm and intonation accuracy) (.56–.58). This pattern is different from the observations when comparing High and Intermediate comprehensibility levels.

Fluency exhibited a similar degree of effect size (.50) to lexicogrammatical features (.42–.53). This contrasts with the findings of the comparison between High and

Intermediate levels. Furthermore, all four features (except for lexical richness) displayed equal effect sizes among the four lexical and grammatical features. This pattern mirrors the findings when comparing High and Intermediate comprehensibility levels.

4.5.2 Objective Features

Table 29 displayed descriptive statistics for 11 objective linguistic features at High, Intermediate, and Low comprehensibility levels. The descriptive statistics revealed that the mean scores for all linguistic features generally demonstrate improvement with increasing comprehensibility levels. Notably, the word stress error ratio exhibited an opposing trend to the overall pattern, as it increased with higher comprehensibility levels. It is also noteworthy that several features displayed substantial dispersion and overlap with adjacent comprehensibility levels. This suggested that certain linguistic features may not differentiate between particular comprehensibility levels.

Table 29

Descriptive Statistics for the 11 Objective Linguistic Features at High, Mid, and Low Comprehensibility Levels

Variable	Level	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Segmental error ratio	High	0.10	0.06	0.09	0.02	0.24	0.95	0.02	0.02
	Mid	0.19	0.08	0.20	0.06	0.41	0.62	−0.09	0.02
	Low	0.35	0.07	0.36	0.20	0.45	−0.74	0.08	0.02
Syllable structure error ratio	High	0.01	0.02	0.00	0.00	0.07	2.36	4.15	0.01
	Mid	0.03	0.02	0.02	0.00	0.09	0.70	−0.27	0.00
	Low	0.08	0.07	0.06	0.00	0.24	1.02	0.02	0.02
Word stress error ratio	High	0.33	0.20	0.35	0.08	0.67	0.46	−1.05	0.06
	Mid	0.19	0.21	0.13	0.00	0.80	1.17	0.59	0.04
	Low	0.15	0.20	0.09	0.00	0.60	1.17	0.16	0.07
Vowel reduction error ratio	High	0.07	0.10	0.03	0.00	0.30	1.44	0.58	0.03
	Mid	0.20	0.17	0.18	0.00	0.71	1.16	1.28	0.03
	Low	0.53	0.15	0.54	0.27	0.72	−0.51	−1.23	0.05
Intonation error ratio	High	0.36	0.18	0.29	0.17	0.75	0.93	−0.38	0.05
	Mid	0.52	0.30	0.50	0.00	1.00	−0.28	−1.06	0.06
	Low	0.57	0.45	0.67	0.00	1.00	−0.35	−1.85	0.15
Articulation rate	High	3.37	0.47	3.39	2.65	4.24	0.17	−1.15	0.14
	Mid	3.01	0.38	3.01	2.42	3.76	0.22	−0.83	0.08
	Low	2.95	0.38	2.82	2.53	3.54	0.27	−1.72	0.13
Lexical error ratio	High	0.04	0.03	0.04	0.02	0.10	0.57	−1.13	0.01
	Mid	0.07	0.05	0.07	0.00	0.19	0.39	−0.96	0.01
	Low	0.07	0.04	0.08	0.00	0.13	−0.36	−1.28	0.01
MTLD	High	38.43	15.26	32.83	17.61	71.68	0.76	−0.44	4.60
	Mid	30.78	12.38	29.62	14.00	53.82	0.36	−1.16	2.48
	Low	28.18	10.73	24.90	20.00	53.82	1.41	0.73	3.58
Lambda	High	0.65	0.41	0.75	0.00	1.40	0.12	−1.20	0.12
	Mid	0.32	0.32	0.33	0.00	1.00	0.52	−0.84	0.06
	Low	0.31	0.43	0.00	0.00	1.00	0.75	−1.33	0.14
Grammatical error ratio	High	0.11	0.04	0.10	0.07	0.19	0.79	−0.78	0.01
	Mid	0.13	0.07	0.14	0.00	0.26	0.00	−1.28	0.01
	Low	0.16	0.08	0.15	0.00	0.25	−0.68	−0.61	0.03
Mean length of AS-unit	High	8.35	1.72	8.50	5.80	11.00	−0.05	−1.44	0.52
	Mid	7.57	1.53	7.75	5.11	11.67	0.46	0.33	0.31
	Low	7.29	1.47	6.67	5.00	10.00	0.29	−0.98	0.49

Subsequently, a robust MANOVA was performed, revealing a significant main effect of comprehensibility on the linguistic features ($H(22) = 43.66$, $p = .004$). Following the initial analysis, 11 sets of Kruskal-Wallis tests were conducted. In contrast to the findings from subjective linguistic features, the results of objective linguistic features revealed effects only on three pronunciation features: segmental error ratio, $H(2) = 22.47$, $p < .001$, $r = .65$, 95% CI [.44, .79]; syllable structure error ratio, $H(2) = 13.99$, $p < .001$, $r = .49$, 95% CI [.23, .69]; and vowel reduction error ratio, $H(2) = 20.78$, $p < .001$, $r = .62$, 95% CI [.40, .77]. The effect sizes ranged from medium to large (.49–.65) (Plonsky & Oswald, 2014). Conversely, the remaining eight features showed no significant differences across comprehensibility levels: word stress error ratio, $H(2) = 5.74$, $p = .06$, $r = .28$, 95% CI [−.01, .53]; intonation error ratio, $H(2) = 3.29$, $p = .19$, $r = .19$, 95% CI [−.11, .46]; articulation rate, $H(2) = 5.09$, $p = .08$, $r = .26$, 95% CI [−.03, .52]; lexical error ratio, $H(2) = 2.97$, $p = .23$, $r = .18$, 95% CI [−.12, .45]; MTLD, $H(2) = 3.50$, $p = .17$, $r = .20$, 95% CI [−.10, .47]; lambda, $H(2) = 5.39$, $p = .07$, $r = .27$, 95% CI [−.02, .52]; grammatical error ratio, $H(2) = 4.20$, $p = .12$, $r = .23$, 95% CI [−.07, .49]; mean length of AS-unit, $H(2) = 2.44$, $p = .30$, $r = .16$, 95% CI [−.14, .43].

Finally, Mann-Whitney's *U* tests were conducted for the three significant linguistic features: segmental error ratio, syllable structure error ratio, and vowel reduction error ratio to compare the linguistic scores between High and Intermediate groups, Intermediate and Low groups, and High and Low groups. Tables 30 through 32 present the test statistics (*Z*), significance test results, effect sizes (*r*), and their corresponding 95% confidence intervals. The results demonstrated significant differences in all three linguistic features between all pairs, with effect sizes (Figure 6) ranging from medium to large (Plonsky & Oswald, 2014).

In the comparison between High and Intermediate groups, the effect sizes of all three features were similar (medium)—segmental error (.49), syllable structure error (.43), and vowel reduction error (.41). When comparing Intermediate and Low groups, segmental and vowel reduction errors exhibited large effect sizes of .63. In contrast, syllable structure error displayed a medium effect size of .43. These findings closely align with the results of subjective linguistic features. Specifically, smaller pronunciation units, such as segmental errors, are crucial in differentiating Intermediate and Low groups but have less power distinguishing High and Intermediate groups. Notably, vowel reduction error—a feature associated with rhythm—exhibited a contrasting trend in the objective linguistic features. Specifically, vowel reduction error

showed a medium effect size of .41 when comparing High and Intermediate groups, while a large effect size of .63 was obtained when comparing Intermediate and Low groups.

Table 30

Results from Mann-Whitney's U Tests for the Three Objective Pronunciation Features of the High and the Intermediate Groups

Variable	*Z*	*p*	*r*	95% CI
Segmental error ratio	−2.94	.002	.49	.19, .70
Syllable structure error ratio	−2.56	.009	.43	.11, .66
Vowel reduction error ratio	−2.48	.012	.41	.10, .65

Note. α = .016

Table 31

Results from Mann-Whitney's U Tests for the Three Objective Pronunciation Features of the Intermediate and the Low Groups

Variable	*Z*	*p*	*r*	95% CI
Segmental error ratio	3.65	< .001	.63	.37, .80
Syllable structure error ratio	2.50	.011	.43	.11, .67
Vowel reduction error ratio	3.65	< .001	.63	.37, .80

Note. α = .016

Table 32

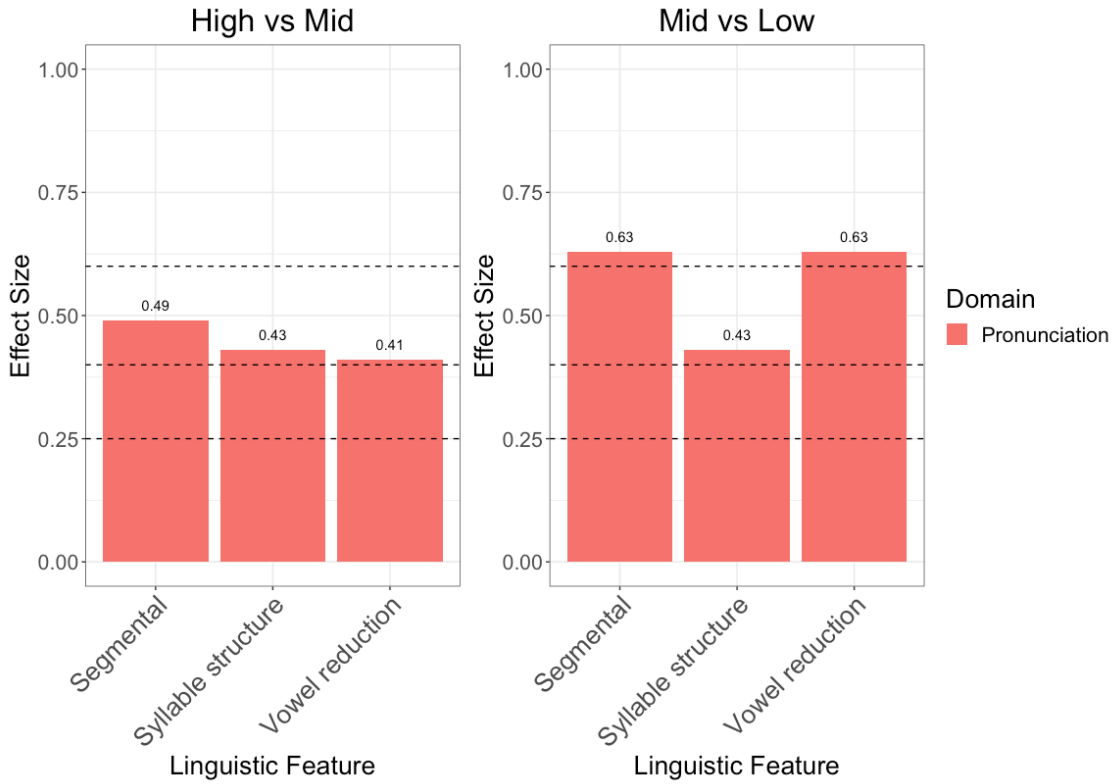
Results from Mann-Whitney's U Tests for the Three Objective Pronunciation Features of the High and the Low Groups

Variable	*Z*	*p*	*r*	95% CI
Segmental error ratio	−3.68	< .001	.82	.60, .93
Syllable structure error ratio	−3.17	< .001	.71	.39, .88
Vowel reduction error ratio	−3.69	< .001	.82	.60, .93

Note. α = .016

Figure 6

Summary of Effect Sizes of Three Objective Pronunciation Features for Comparisons Between High and Intermediate Groups, and Intermediate and Low Groups



Note. The dashed lines indicate reference points corresponding to effect sizes for small (.25), medium (.40), and large (.60) based on Plonsky and Oswald (2014).

Chapter 5 Discussion

This study explored how linguistic features within the domains of pronunciation, fluency, vocabulary, and grammar differentially influence comprehensibility, comparing subjective and objective linguistic assessments. While the analyses of subjective and objective linguistic features yielded divergent outcomes depending on the statistical analyses used, pronunciation and fluency, overall, had a stronger influence on comprehensibility than lexicogrammar. The subsequent paragraphs present separate discussions based on correlation and multiple regression analyses and MANOVA.

5.1 Correlations Between Subjective and Objective Linguistic Features

Significant, but varying degrees of, associations were found between five subjective linguistic features and their objective counterparts: segmental error (.78, with segmental error ratio, and .54 with syllable structure error ratio), rhythm error (.73, with vowel reduction error ratio), speech rate (−.33, with articulation rate), lexical richness (−.36, with MTLD, and −.42 with lambda), and grammatical error (.36, with grammatical error ratio) (see Tables 1 and 15). Saito et al. (2017) also found significant correlations between these five subjective features and their objective counterparts. When the correlation coefficients were compared between the previous study and the current study, three of the features exhibited similar levels of correlation: segmental

error (.64 in Saito et al., 2017 and .78 in the present study), rhythm error (.74 and .73), and speech rate (.43 and −.33). However, there were divergent correlations in lexical richness (.74 and −.36) and grammatical accuracy (.67 and .36). Although Saito et al. (2017) also found significant correlations for word stress error, intonation error, lexical appropriateness, and grammatical complexity, the current study did not find significant correlations for these features. These varying degrees of correlations and inconsistent results between this study and Saito et al (2017) indicate that subjective and objective linguistic assessments tap into somewhat different constructs even though same/similar labels are assigned to the scores.

5.2 Hierarchical Multiple Regression Analysis

The hierarchical multiple regression analyses produced consistent results for subjective and objective linguistic features. Specifically, pronunciation and fluency demonstrated a more substantial influence on comprehensibility than lexicogrammar. The results align with previous findings (Crowther et al., 2015; Saito & Shintani, 2015; Saito et al., 2016; 2017).

Nevertheless, it is notable that the relative importance ratio of pronunciation to lexicogrammar in relation to comprehensibility in the current study differed from prior findings (Saito et al., 2016; 2017). In this study, irrespective of subjective or objective

linguistic features, pronunciation had almost three times more impact on comprehensibility than did lexicogrammar—although this could vary depending on the order of variable entry. In contrast, Saito et al. (2016) reported that pronunciation had 1.6 times more impact on comprehensibility than lexicogrammar (a combination of lexicogrammar accuracy and sophistication). Similarly, Saito et al. (2017) found that pronunciation had 1.3 times more impact than lexicogrammar.

This heightened impact of pronunciation relative to lexicogrammar compared with the previous findings might be attributed to the fact that pronunciation errors were more noticeable in the speakers in this study compared to those in previous studies. This was likely due to lower levels of pronunciation proficiency stemming from limited exposure to English-speaking environments in the current participants.

Prior research has illustrated that increased exposure to the native input of the target language leads to improved pronunciation proficiency, including segmental and suprasegmental features, even in the absence of explicit pronunciation instruction (e.g., Munro & Derwing, 2008; Trofimovich & Baker, 2006). For instance, Munro and Derwing (2008) explored the longitudinal improvement of 10 English vowels among Mandarin and Slavic English language learners residing in target language environments for one year, and reported a significant enhancement in vowel production.

Similarly, Trofimovich and Baker (2006), investigating the effect of length of residence in a target language environment on suprasegmental accuracy for Korean learners of English in the United States, demonstrated a significant correlation between length of residence and holistic accuracy in suprasegmental features, including stress, rhythm, and intonation.

In the current study, 28 out of 45 speakers had experience living in English-speaking countries. However, the median length of their residence in these countries was approximately four months. In contrast, all the speakers in Saito et al. (2016) were residents of Canada, with roughly 80% of them having spent between one and 41 years in the country. Similarly, the speakers in Saito et al. (2017) were also residents of Canada. Although they resided in a French-speaking environment, they likely had more exposure to native English input than the participants in this study.

5.3 MANOVA

With regard to which specific linguistic features differentiate between different comprehensibility levels (MANOVA), the results showed disparities between the subjective and objective linguistic features. In the subjective linguistic analyses, all nine features differentiated across all three comprehensibility levels. In contrast, in the objective linguistic features, only three pronunciation features (the segmental error ratio,

syllable structure error ratio, and vowel reduction error ratio) distinguished speakers across all three comprehensibility levels. The following sections provide separate discussions of the results from subjective and objective linguistic features.

5.3.1 Subjective Features

5.3.1.1 High-to-Intermediate Comparison

First, the outcomes of subjective linguistic features are discussed, focusing on the comparison between High and Intermediate groups. In this comparison, the impact of pronunciation and fluency on the group difference was stronger than those of lexical and grammatical features. These findings align with the results of the correlation and regression analyses. This heightened influence of pronunciation and fluency features on comprehensibility can be understood by considering the listening process.

Rost (2005) categorized the listening process into three phases—decoding, comprehension, and interpretation. There are four sub-phases within the decoding phase—attention, perception, word recognition, and syntactic parsing. After attention is directed to incoming speech, perception primarily helps the listener to make sense of the speech signal. Consequently, this phase is closely associated with processing speech sounds, including aspects such as pronunciation and fluency. Subsequently, in the word recognition phase, the listener identifies words and promptly activates their lexical

knowledge linked to the recognized words. In this phase, lexical information is processed, encompassing factors like accuracy and complexity. Finally, the syntactic parsing phase involves processing the language to derive meaning, necessitating a syntactic mapping of incoming speech onto a grammatical model. Therefore, this phase is linked to the processing of grammatical features, encompassing features like accuracy and complexity.

In the sequence of these phases, pronunciation and fluency are the initial cues for assessing comprehensibility, followed by lexical and grammatical features. The relative impacts found for pronunciation, fluency, vocabulary, and grammar on comprehensibility in this study suggest that features necessary for the earlier (speech) perception process is primarily more important for improving comprehensibility levels.

Notably, among the four aspects of pronunciation, suprasegmental accuracy exhibited a greater power in distinguishing High and Intermediate comprehensibility levels than segmental accuracy. The presence of a hierarchical pattern within the three suprasegmental features is particularly intriguing, suggesting that as the units of pronunciation increase in size, their influence on comprehensibility becomes more pronounced. For instance, in the comparison between word stress and rhythm errors, the latter (a larger unit of pronunciation) has a more substantial influence on the difference

in comprehensibility between High and Intermediate groups. Similarly, when comparing rhythm errors with intonation errors, the latter (a larger unit of pronunciation) demonstrates a greater impact on the difference in comprehensibility between High and Intermediate groups. Additionally, an appropriate speech rate is as crucial as accurate pronunciation to differentiate between High and Intermediate groups.

A comparison of the relative effects of lexical and grammatical features shows that all four features of accuracy and complexity equally differentiated the comprehensibility levels. These findings indicate that the correct utilization of lexical items and grammatical structure is as essential as the sophisticated use of lexical items and grammatical structures—such as employing varied and less common vocabulary and complex sentence structures.

5.3.1.2 Intermediate-to-Low Comparison

Turning to the Intermediate-to-Low group comparison, again the effect of pronunciation was stronger on the differentiation between the groups than lexical and grammatical features. Among the sound-related variables, in contrast to the comparison between High and Intermediate groups, the effect of segmental accuracy was generally stronger than that of suprasegmental accuracy, although word stress errors (a

suprasegmental feature) displayed a slightly stronger power in differentiating speakers between Intermediate and Low comprehensibility levels than segmental accuracy.

Furthermore, the impact of fluency on comprehensibility displayed varying trends. Fluency had less power in differentiating between Intermediate and Low groups than between the High and Intermediate groups.

These results revealed that the relatively smaller components of pronunciation (segmental and word stress accuracy) displayed a stronger capability in differentiating between Intermediate and Low comprehensibility levels than the larger sound features (rhythm accuracy, intonation accuracy, and speech rate).

The comparison between lexical and grammatical features revealed that both features differentiated speakers between Intermediate and Low groups (to a similar degree). This pattern aligns with the observations made in High-Intermediate comparison, indicating that for all levels both lexical and grammatical features hold equal importance.

An interesting trend emerged when the effect sizes of two lexical features were compared. Lexical appropriateness had a greater power in differentiating between Intermediate and Low groups than lexical richness. This suggests that, for speakers at Intermediate and Low comprehensibility levels, it is more essential to employ

appropriate lexical items that suit the context rather than utilizing less-frequent lexical items.

5.3.2 Objective Features

5.3.2.1 High-to-Intermediate Comparison

Next, the results of objective linguistic features will be explored, first focusing on High-to-Intermediate group comparison. Consistent with the results of the subjective linguistic features, pronunciation features (segmental, syllable structure, and vowel reduction) possessed a greater power to distinguish speakers between High and Intermediate levels than lexical and grammatical features. Nevertheless, it is noteworthy that the results appear quite extreme. That is, none of the lexical and grammatical features yielded significant results, nor did fluency.

Notably, among the three pronunciation features that displayed significant results, the smallest unit (segmental accuracy) exhibited the greatest discriminatory power, followed by syllable structure and vowel reduction. These outcomes contrast with the results of subjective linguistic assessments, which indicated that larger units of pronunciation (vowel reduction accuracy) demonstrated a stronger discriminatory power compared to smaller units (segmental accuracy).

5.3.2.2 Intermediate-to-Low Comparison

Finally, we discuss Intermediate-and-Low group comparison. Again, three pronunciation features (segmental error, syllable structure error, and vowel reduction error) differentiated the speakers between these levels, while none of the fluency, lexical, and grammatical features yielded significant results. Among the three pronunciation features, the smallest pronunciation unit (segmental accuracy) exhibited the most substantial discriminatory power. This aligns with the findings from subjective linguistic features. Moreover, vowel reduction accuracy played a comparable role to segmental accuracy in distinguishing speakers at this level of comprehensibility.

Additionally, when comparing the effect sizes of the three pronunciation features between the Intermediate-to-Low group and the High-to-Intermediate group, it was observed that the former exhibited larger sizes. This substantiates the findings from the MANOVA on subjective linguistic features and supports the Rost's (2005) listening process, implying that the pronunciation accuracy holds greater importance for speakers with lower comprehensibility levels.

Chapter 6 Conclusion

The present study concludes by summarizing the findings, drawing pedagogical implications, recognizing limitations, and suggesting directions for future research.

6.1 Summary

This study examined the relationship between comprehensibility and various linguistic features, encompassing pronunciation, fluency, vocabulary, and grammar. Additionally, it investigated whether these effects varied depending on the subjectivity of linguistic assessments. The findings yielded valuable insights. First, consistent with prior research, features of speech sounds—such as pronunciation accuracy and speech rate—exhibited stronger associations with comprehensibility compared to lexical and grammatical accuracy and complexity. This pattern remained consistent regardless of the type of linguistic assessments used.

Second, the influence of specific linguistic features varied in accordance with the comprehensibility levels of the speakers, which is a novel finding. To elaborate, while pronunciation accuracy and fluency features held greater importance than lexical and grammatical features for speakers across all comprehensibility levels, this trend was particularly pronounced among speakers with lower comprehensibility levels. Moreover, for speakers falling within Low to Intermediate comprehensibility range, the

precision of smaller speech units—such as segmental and word stress accuracy—outweighed larger units like rhythm and intonation accuracy. Lexical and grammatical accuracy and complexity played a comparatively smaller role in this context.

Conversely, speakers with Intermediate to High comprehensibility levels placed greater emphasis on larger speech units, including rhythm and intonation accuracy, and speech rate, in determining comprehensibility. Lexical and grammatical accuracy and complexity were of similar importance in this group.

6.2 Pedagogical Implications

Based on the findings, the present study carries several pedagogical implications. First, pronunciation instruction—focusing on segmentals, syllable structure, and rhythm—is essential for all L2 speakers, regardless of their comprehensibility levels. These features consistently emerged as significant factors related to comprehensibility in both subjective and objective linguistic assessments. Indeed, explicit pronunciation instruction has been shown to enhance learners' pronunciation accuracy, encompassing both segmental and suprasegmental aspects (e.g., Saito, 2013; Saito & Saito, 2017).

As discussed earlier, the relative importance of specific linguistic features for comprehensibility varies depending on the comprehensibility levels. Ideally, speakers should be grouped into different classes based on their proficiency levels. If this can be

achieved, a tailored teaching approach can be adopted. Specifically, for L2 speakers with Intermediate to High comprehensibility levels, teachers should prioritize the accurate production of rhythm and intonation over segmental and word stress accuracy. Additionally, these learners should practice speaking at a faster rate. Simultaneously, they should work on producing accurate and sophisticated vocabulary and grammar, including a variety of less frequent words and more complex sentence structures.

Conversely, for L2 speakers with Intermediate to Low comprehensibility levels, the focus should be on teaching segmental and word stress accuracy over rhythm and intonation accuracy. Speaking at a faster speech rate is also important at this level but less critical compared to speakers at higher levels. Once proficiency in pronunciation is achieved, instruction should shift towards accurate use of vocabulary and grammar, followed by the incorporation of varied and less frequent words and more complex sentence structures in later stages.

However, it is often practically challenging to divide students into multiple classes based on their proficiency levels, and the available class time for pronunciation instruction is typically limited. Additionally, teachers may not always possess the necessary expertise to teach specific pronunciation features, as effective pronunciation instruction often requires advanced techniques, such as demonstrating subtle differences

in tongue movements for segmental teaching or variations in reduced and unreduced vowels for rhythm instruction. In such situations, for example, integrating shadowing practice into their classes might be advisable. Foote and McDonough (2017), who engaged L2 learners in shadowing short dialogues for a minimum of ten minutes at least four times per week over eight weeks, demonstrated a significant overall improvement in their comprehensibility.

6.3 Limitations

There are some potential limitations in the present study. The first limitation concerns fluency features. In the current study, only speech rate (i.e., speed fluency) was considered, as a feature of fluency. However, fluency can be further subdivided into three dimensions: speed, breakdown, and repair fluency (Tavakoli & Skehan, 2005). Previous studies utilizing objective linguistic assessments have indicated that all these sub-dimensions are associated with comprehensibility (e.g., Suzuki & Kormos, 2020; Trofimovich & Isaacs, 2012). This study intentionally chose not to include features of breakdown and repair fluency because the 1000-point sliding scales used in this research have not been validated for effectively capturing these two features (Saito et al., 2017). Interestingly, Bosker et al. (2013) conducted a study in which they had native Dutch raters subjectively evaluate speed, breakdown, and repair fluency of L2 speech

using 9-point scales. Their findings demonstrated that these subjective assessments were statistically predicted by their objective counterparts. Given these findings, future research is encouraged to investigate whether the 1000-point sliding scales can also adequately measure breakdown and repair fluency, and to explore the associations between these features and comprehensibility.

The second limitation concerns the statistical methodologies employed to determine the relative importance of linguistic features in relation to comprehensibility. In the current study, hierarchical multiple regression analyses were utilized. However, as documented by Keith (2019), this approach has a potential limitation: the order in which variables are entered can greatly impact the resulting R^2 values. To overcome this challenge, Mizumoto (2023) suggests the use of dominance analysis, also known as Shapley value regression. This statistical technique decomposes the overall R^2 values and allocates these effects to each of the significant independent variables (Lai et al., 2022). It allows for a more nuanced understanding of how collinear independent variables contribute to the dependent variable. This analysis can be conducted using analytical tools such as RStudio and MATLAB software. Future research is encouraged to employ such a technique to reevaluate the relative importance of various linguistic features in relation to comprehensibility.

Finally, the use of brief speech excerpts may have posed limitations on the ability to conduct adequate linguistic assessments. This study closely followed the methodology of previous studies (e.g., Saito et al., 2017), which employed only 30-second short speech samples for the analysis of various linguistic features. If future research is to employ speech segments for linguistic analysis, it may be advisable to extract speech samples based on word count rather than on their time duration. Alternatively, it could be suggested that future research endeavors to consider analyzing complete speech samples to attain more robust linguistic evaluations because determining a minimum word count threshold that ensures a comprehensive assessment of linguistic features is a complex task. The limited associations between specific subjective and objective linguistic features found in this study may change if these features are assessed with a larger amount of speech samples. Moreover, it allows for the examination of discourse features, which could not be included in the present study.

6.4 Future Directions

Several research directions hold promise for future investigation. First, as our understanding of the linguistic aspects contributing to comprehensibility advances, it is crucial to investigate the effectiveness of pedagogical interventions aimed at enhancing comprehensibility. Pronunciation instruction has been a focal point in numerous studies,

with promising outcomes for comprehensibility development; however, to the best of the author's knowledge, there has not been an examination of the relationship between the progress in fluency, vocabulary, and grammatical features and the enhancement of comprehensibility. For the effectiveness of pronunciation instruction, Saito and Saito (2017) delved into the impact of form-focused instruction on the development of suprasegmental skills among Japanese EFL learners at the beginner level. Their findings revealed noteworthy improvements in various suprasegmental facets, including word stress, rhythm, intonation, as well as overall comprehensibility. Likewise, Zhang and Yuan (2020) administered both segmental- and suprasegmental-based instruction to Chinese learners of English over an 18-week duration. Their research indicated that both types of instruction effectively enhanced comprehensibility, notably in read-aloud tasks. The group receiving suprasegmental instruction also displayed improved comprehensibility in spontaneous speech contexts, with these gains remaining consistent during a delayed post-test. Furthermore, Derwing et al. (2014) focused on immigrant L2 English learners with substantial experience in an English-speaking environment. Their study demonstrated the efficacy of explicit pronunciation instruction in enhancing the comprehensibility of L2 learners, even when their pronunciation had become "fossilized" after prolonged residence in the target language environment. As

previously noted, it remains unexplored whether the advancement of various linguistic aspects other than pronunciation, including fluency, vocabulary, and grammar, contributes to the enhancement of comprehensibility. For instance, task-repetition exercises conducted over a brief three-day period have demonstrated their effectiveness in enhancing specific fluency components, such as speed and breakdown (Suzuki, 2021). Likewise, both input-based and output-based activities within the classroom setting have proven effective in enhancing productive vocabulary skills (Teng & Zhang, 2021). Nevertheless, these intervention studies have not investigated whether these linguistic advancements lead to an improvement in comprehensibility. Therefore, it would be valuable for future research to further examine whether these language enhancements result in improved comprehensibility.

The final consideration for potential future research relates to the methodology used to gauge comprehensibility. Comprehensibility is defined as the *perception* by listeners regarding the ease or difficulty of understanding L2 speech. Accordingly, prior studies have primarily relied on subjective assessments, employing 9-point Likert-type scales (Trofimovich & Isaacs, 2012; Suzuki & Kormos, 2020) and 1000-point sliding scales (Saito et al., 2017) to evaluate comprehensibility. Subjective ratings offer the advantage of simplicity and ease of application, even in natural settings. However, a

drawback of these ratings is the instability of an individual's framework of reference, which can change over time due to adaptation processes or in response to motivational and emotional shifts (Schnotz & Kürschner, 2007). In light of these challenges, the field of educational psychology has developed various objective methodologies to measure cognitive effort (Schnotz & Kürschner, 2007). In physiological approaches, examples include the utilization of galvanic skin response, pupillary dilation, and heart rate variability. Another objective measure is the performance-based method, also known as the dual-task methodology, where participants engage in a primary task while simultaneously performing a secondary task, typically a simple reaction task. The performance on the secondary task can serve as an indicator of the cognitive effort imposed by the primary task (Brünken et al., 2003). In the realm of comprehensibility research, a few studies have employed such objective methodologies to quantify comprehensibility. Hahn (2004) utilized the aforementioned dual-task methodology, while Munro and Derwing (1995b) and Ludwig and Mora (2017) employed reaction time as a measure of comprehensibility. Nevertheless, these studies did not explore the linguistic features influencing these objective measures. A promising avenue for future research would be to employ these objective measures of comprehensibility and investigate their linguistic correlates.

References

Abrahamsson, N. (2012). Age of onset and nativelike L2 ultimate attainment of morphosyntactic and phonetic intuition. *Studies in Second Language Acquisition*, *34*(2), 187–214. https://doi.org/10.1017/S0272263112000022

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, *59*(2), 249–306. https://doi.org/10.1111/j.1467-9922.2009.00507.x

Albrechtsen, D., Henriksen, B., & Færch, C. (1980). Native speaker reactions to learners' spoken interlanguage. *Language Learning*, *30*(2), 365–396. https://doi.org/10.1111/j.1467-1770.1980.tb00324.x

Anderson–Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, *42*(4), 529–555 https://doi.org/10.1111/j.1467-1770.1992.tb01043.x

Bartlett, M. S. (1954). A note on the multiplying factors for various χ^2 approximations. *Journal of the Royal Statistical Society (Series B)*, *16*(2), 296–298. https://doi.org/10.1111/j.2517-6161.1954.tb00174.x

Birdsong, D. (2006). Age and second language acquisition and processing: A selective overview. *Language Learning*, *56*(1), 9–49. https://doi.org/10.1111/j.1467-9922.2006.00353.x

Boersma, P., & Weenink, D. (2016). *Praat: Doing phonetics by computer* (Version 6.0.16) [Computer software]. http://www.praat.org/

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175. https://doi.org/10.1177/0265532212455394

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* The Guilford Press.

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 53–61. https://doi.org/10.1207/S15326985EP3801_7

Bulut, H. (2019). An R Package for Multivariate Hypothesis Tests: MVTests. *Technological Applied Sciences*, *14*(4), 132–138. https://dergipark.org.tr/tr/pub/nwsatecapsci/issue/49784/599944

Choi, K., & Marden, J. (1997). An approach to multivariate rank tests in multivariate analysis of variance. *Journal of the American Statistical Association*, *92*(440), 1581–1590. https://doi.org/10.1080/01621459.1997.10473680

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *The Modern Language Journal*, *99*(1), 80–95. https://doi.org/10.1111/modl.12185

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*, *49*(4), 814–837. https://doi.org/10.1002/tesq.203

De Clercq, B., & Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: Syntactic elaboration and diversity. *The Modern Language Journal*, *101*(2), 315–334. https://doi.org/10.1111/modl.12396

Derwing, T. M. (2003). What do ESL students say about their accents? *Canadian Modern Language Review*, *59*(4), 547–567. https://doi.org/10.3138/cmlr.59.4.547

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(1), 1–16. https://doi.org/10.1017/S0272263197001010

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

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*(4), 533–557. https://doi.org/10.1017/S0272263109990015

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning*, *54*(4), 655–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Deterding, D. (2001). The measurement of rhythm: A comparison of Singapore and British English. *Journal of Phonetics*, *29*(2), 217–230. https://doi.org/10.1006/jpho.2001.0138

Educational Testing Service. (2008). *Linking English-language test scores onto the Common European Framework of Reference: An application of standard-setting methodology.* https://www.ets.org/Media/Research/pdf/RR-08-34.pdf

Educational Testing Service. (2023). *TOEFL iBT® integrated speaking rubric.* https://www.ets.org/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf

Educational Testing Service. (2023). *TOEFL iBT® independent speaking rubric.* https://www.ets.org/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, *37*(3), 313–326. https://doi.org/10.1111/j.1467-1770.1987.tb00573.x

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, *39*(3), 399–423. https://doi.org/10.2307/3588487

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. SAGE Publications Ltd.

Flege, J. E., Bohn, O-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, *25*(4), 437–470. https://doi.org/10.1006/jpho.1997.0052

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, *97*(5), 3125–3134. https://doi.org/10.1121/1.413041

Foote, J. A., & McDonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation*, *3*(1), 34–56. https://doi.org/10.1075/jslp.3.1.02foo

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. https://doi.org/10.1093/applin/21.3.354

Gamer, M., Lemon, J., & Singh, I. F. P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement* (Version 0.84.1) [Computer software]. https://CRAN.R-project.org/package=irr

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, *19*(2), 149–161. https://doi.org/10.1007/BF02289162

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*(2), 201–223. https://doi.org/10.2307/3588378

Hothorn T., Hornik, K., van de Wiel, M. A., Zeileis, A. (2006). A Lego system for conditional inference. The American Statistician, 60(3), 257–263. https://doi.org/10.1198/000313006X118430

International Development Program (IDP) Education. (2023). *IELTS speaking band descriptors.* https://assets.ctfassets.net/unrdeg6se4ke/4HClJPN2BGdO1fcc018Gz9/f5e625eb26d075a4d8b5151da0b90709/Speaking-Band-descriptors.pdf

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*(3), 475–505. https://doi.org/10.1017/S0272263112000150

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49. https://doi.org/10.1093/applin/amm017

Jarek, S. (2012). *mvnormtest. Normality test for multivariate variables* (Version 0.1.9) [Computer software]. https://CRAN.R-project.org/package=mvnormtest

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Kaiser, H. F. (1970). A second-generation Little Jiffy. *Psychometrika*, *35*(4), 401–415. https://doi.org/10.1007/BF02291817

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115

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*(4), 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

Khatib, M., & Farahanynia, M. (2020). Planning conditions (strategic planning, task repetition, and joint planning), cognitive task complexity, and task type: Effects on L2 oral performance. *System*, *93*, 1–12. https://doi.org/10.1016/j.system.2020.102297

Keith, T. Z. (2019). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling* (3rd ed.). Routledge. https://doi.org/10.4324/9781315162348

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*(4), 554–564. https://doi.org/10.1016/j.system.2012.10.012

Lai, J., Zou, Y., Zhang, S., Zhang, X., & Mao, L. (2022). glmm.hp: An R package for computing individual effect of predictors in generalized linear models. *Journal of Plant Ecology*, *15*(6), 1302–1307. https://doi.org/10.1093/jpe/rtac096

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. Routledge.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

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

Ludwig, A., & Mora, J. C. (2017). Processing time and comprehensibility judgments in non-native listeners' perception of L2 speech. *Journal of Second Language Pronunciation*, *3*(2), 167–198. https://doi.org/10.1075/jslp.3.2.01lud

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488. https://doi.org/10.1177/0265532207080767

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Meara, P. (2018). *P_Lex_lognostics* (Version 2.31) [Computer software]. https://www.lognostics.co.uk/tools/P_Lex/P_Lex.htm

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16*(3), 5–19.

Meara, P., & Miralpeix, I. (2016). *Tools for researching vocabulary*. Multilingual Matters. https://doi.org/10.21832/9781783096473

Michel, M. (2017). Complexity, accuracy, and fluency in L2 production. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 50–68). Routledge. https://doi.org/10.4324/9781315676968

Mikami, R. (2019). Phonological and lexicogrammatical factors underlying Comprehensibility: A study investigating the effect of listeners' first language backgrounds. *Language Education & Technology*, *56*, 23–50.

Mizumoto, A (2023). Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning*, *73*(1), 161–196. https://doi.org/10.1111/lang.12518

Mora, C. J., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly*, *46*(4), 610–641. https://doi.org/10.1002/tesq.34

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

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*(3), 289–306. https://doi.org/10.1177/002383099503800305

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

Munro, M. J., & Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning*, *58*(3), 479–502. https://doi.org/10.1111/j.1467-9922.2008.00448.x

Munro, M., & Mann, V. (2005). Age of immersion as a predictor of foreign accent. *Applied Psycholinguistics*, *26*(3), 311–344. https://doi.org/10.1017/S0142716405050198

Pang, F., & Skehan, P. (2014). Self-reported planning behavior and second language performance on narrative retelling. In P. Skehan (Ed.), *Processing perspective on task performance* (pp. 95–127). John Benjamins. https://doi.org/10.1075/tblt.5

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R^2 values. *The Modern Language Journal*, *102*(4), 713–731. https://doi.org/10.1111/modl.12509

Plonsky, L., & Oswald, F. (2014). How big is "Big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

R Core Team (2020). *R: A language and environment for statistical computing* (Version 1.3.1093) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2020). *psych: Procedures for Psychological, Psychometric, and Personality Research* (Version 2.0.12) [Computer software]. https://CRAN.R-project.org/package=psych

Rost, M. (2005). L2 listening. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 503–527). Lawrence Erlbaum.

Saito, K. (2013). The acquisitional value of recasts in instructed second language speech learning: Teaching the perception and production of English /ɹ/ to adult Japanese learners. *Language Learning*, *63*(3), 499–529. https://doi.org/10.1111/lang.12015

Saito, K., & Akiyama, Y. (2017). Linguistic correlates of second language Japanese speech. *Journal of Second Language Pronunciation*, *3*(2), 199–217. https://doi.org/10.1075/jslp.3.2.02sai

Saito, Y., & Saito, K. (2017). Differential effects of instruction on the development of second language comprehensibility, word stress, rhythm, and intonation: The case of inexperience Japanese EFL learners. *Language Teaching Research*, *21*(5), 589–608. https://doi.org/10.1177/1362168816643111

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for

learners at different ability levels. *Applied Psycholinguistics*, *37*(2), 217–240. https://doi.org/10.1017/S0142716414000502

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*(4), 439–462. https://doi.org/10.1093/applin/amv047

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2015). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition*, *19*(3), 597–609. https://doi.org/10.1017/S1366728915000255

Schnotz, W., & Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review*, *19*(4), 469–508. https://doi.org/10.1007/s10648-007-9053-4

Shirkey, C. E., & Dziuban, D. C. (1976). A note on some sampling characteristics of the measure of sampling adequacy (MSA). *Multivariate Behavioral Research*, *11*(1), 125–128. https://doi.org/10.1207/s15327906mbr1101_9

Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arachchige, C., Arppe, A., Baddeley, A., Barton, K., Bolker, B., Borchers, H. W., Caeiro, F., Champely.,

S., Chessel., D., Chhay, L., Cooper, N., Cummins., C., Dewey, M., Doran, H. C.,...Zeileis, A. (2021). *DescTools: Tools for descriptive statistics* (Version 0.99.44) [Computer software]. https://cran.r-project.org/package=DescTools

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*(1), 1–14. https://doi.org/10.1017/S026144480200188X

Skehan, P., & Tavakoli, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). John Benjamins. https://doi.org/10.1075/lllt.11.15tav

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, *24*(4), 402–433. https://doi.org/10.2307/747605

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9781410604491

Sugimori, M., Sugimori, N., Nakanishi, Y., & Shimizu, Y. (1997). *Onsei eigo no riron to jissen* [Spoken English: Theory and practice]. Eiho-sha.

Suzuki, Y. (2021). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. *Language Learning*, *71*(2), 285–325. https://doi.org/10.1111/lang.12433

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143–167. https://doi.org/10.1017/S0272263119000421

Tabacknick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson Education Limited.

Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics*, *25*(1), 1–24. https://doi.org/10.1006/jpho.1996.0031

Tavakoli, P. (2018). L2 development in an intensive Study Abroad EAP context. *System*, *72*, 62–74. https://doi.org/10.1016/j.system.2017.10.009

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). John Benjamins. https://doi.org/10.1075/lllt.11

Teng, M. F., & Zhang, D. (2021). Task-induced involvement load, vocabulary learning in a foreign language, and their association with metacognition. *Language Teaching Research*, Advance online publication. https://doi.org/10.1177/13621688211008798

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*(1), 1–30. https://doi.org/10.1017/S0272263106060013

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*(4), 905–916. https://doi.org/10.1017/S1366728912000168

Varonis, E. M., & Gass, S. (1982). The comprehensibility of non-native speech. *Studies in Second Language Acquisition*, *4*(2), 114–136. https://doi.org/10.1017/S027226310000437X

Wells, J. C. (2008). *Longman pronunciation dictionary* (3rd ed.). Pearson Education Limited.

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). Academic Press.

Wilcox, R. R., & Schönbrodt, F. D. (2014). *The WRS package for robust statistics in R* (version 0.24). [Computer software] http://r-forge.r-project.org/projects/wrs/

Winters, S., & O'Brien, M. G. (2013). Perceived accentedness and intelligibility: The relative contributions of F0 and duration. *Speech Communication*, *55*(3), 486–507. https://doi.org/10.1016/j.specom.2012.12.006

Yu, H., & Lowie, W. (2020). Dynamic paths of complexity and accuracy in second language speech: A longitudinal case study of Chinese learners. *Applied Linguistics*, *41*(6), 855–877. https://doi.org/10.1093/applin/amz040

Zhang, R., & Yuan, Z. (2020). Examining the effects of explicit pronunciation instruction on the development of L2 pronunciation. *Studies in Second Language Acquisition*, *42*(4), 905–918. https://doi.org/10.1017/S0272263120000121

Appendixes

Appendix 1

Descriptive Statistics for Linguistic Raters' Understanding of the Nine Subjective Linguistic Features

Variable	*M*	*SD*	Median	Min	Max	Skew	Kurtosis	*SE*
Segmental error	8.6	0.89	9.0	7.0	9.0	−1.07	−0.92	0.40
Word stress error	8.6	0.89	9.0	7.0	9.0	−1.07	−0.92	0.40
Intonation error	8.8	0.45	9.0	8.0	9.0	−1.07	−0.92	0.20
Rhythm error	8.2	1.30	9.0	6.0	9.0	−0.82	−1.29	0.58
Speech rate	8.6	0.55	9.0	8.0	9.0	−0.29	−2.25	0.24
Lexical appropriateness	8.8	0.45	9.0	8.0	9.0	−1.07	−0.92	0.20
Lexical richness	8.8	0.45	9.0	8.0	9.0	−1.07	−0.92	0.20
Grammatical accuracy	9.0	0.00	9.0	9.0	9.0	—	—	0.00
Grammatical complexity	8.8	0.45	9.0	8.0	9.0	−1.07	−0.92	0.20

Note. 1 = I did not understand this concept at all, 9 = I understand this concept well.

Appendix 2

Results from a Principal Component Analysis of the Nine Subjective Linguistic Features with 9-factor Solution Followed by No Rotation

Variable	Factor loading								
	1	2	3	4	5	6	7	8	9
Segmental error	**.89**	−.33	.26	.10	−.03	−.09	.12	−.02	.02
Word stress error	**.90**	−.36	.13	.09	.15	.00	−.14	.00	.00
Rhythm error	**.93**	−.33	−.06	.00	−.04	.09	.04	.06	−.09
Intonation error	**.91**	−.32	−.18	−.06	−.04	.14	.01	−.04	.07
Speech rate	**.93**	−.12	−.29	−.09	−.01	−.19	−.02	−.01	−.01
Lexical appropriateness	**.88**	.33	.15	−.26	.13	.02	.05	.01	.00
Lexical richness	**.85**	**.47**	−.13	.17	.05	.00	.03	.09	.03
Grammatical accuracy	**.92**	.27	.17	−.06	−.20	−.01	−.09	.03	.01
Grammatical complexity	**.88**	**.44**	−.04	.12	.01	.04	.00	−.12	−.04

Note. The factors were extracted without rotation. Factor loadings above .40 are in bold.

Appendix 3

Results from a Principal Component Analysis of the Nine Subjective Linguistic Features with Two-factor Solution Followed by No Rotation

Variable	Factor loading	
	1	2
Segmental error	**.89**	−.33
Word stress error	**.90**	−.36
Rhythm error	**.93**	−.33
Intonation error	**.91**	−.32
Speech rate	**.93**	−.12
Lexical appropriateness	**.88**	.33
Lexical richness	**.85**	**.47**
Grammatical accuracy	**.92**	.27
Grammatical complexity	**.88**	**.44**

Note. The factors were extracted without rotation. Factor loadings above .40 are in bold.

Appendix 4

Results from a Principal Component Analysis of the 11 Objective Linguistic Features with 11-factor Solution Followed by No Rotation

Variable	Factor loading										
	1	2	3	4	5	6	7	8	9	10	11
Phonemic substitution ratio	**.91**	.13	−.04	.22	−.12	.00	.04	−.05	.22	.13	.13
Vowel reduction error ratio	**.90**	.23	.09	.27	−.09	−.01	−.05	−.04	.17	.07	−.14
Syllable structure error ratio	**.72**	**.40**	.15	.03	.03	−.12	−.06	.28	−.39	−.24	.02
Intonation error ratio	.14	**−.78**	.17	.32	−.03	.16	.01	.31	−.21	.26	.00
Articulation rate	−.35	**.67**	−.21	.08	.31	−.33	.19	.19	−.06	.31	.00
Mean length of AS-unit	−.31	.12	**.69**	.17	−.12	.05	**.53**	.21	.18	−.12	.00
Lambda	−.29	.29	**−.59**	.29	−.13	**.47**	−.10	.36	.13	−.10	.00
Grammatical error ratio	.23	−.31	**−.51**	.38	.39	−.02	**.48**	−.20	−.08	−.14	−.01
Word stress error ratio	**−.40**	−.18	.11	**.63**	.12	**−.46**	−.37	.05	.13	−.13	.01
Lexical error ratio	.30	−.18	.15	−.32	**.79**	.12	−.13	.23	.21	−.04	.00
MTLD	−.22	**.41**	**.42**	.39	.32	**.46**	−.15	−.31	−.15	.06	.01

Note. The factors were extracted without rotation. Factor loadings above .40 are in bold.

Appendix 5

Results from a Principal Component Analysis of the 11 Objective Linguistic Features With 5-factor Solution Followed by Promax Rotation (Pattern Matrix)

Variable	Factor loading				
	1	2	3	4	5
Vowel reduction error ratio	**1.01**	−.08	.07	.08	−.08
Phonemic substitution ratio	**.95**	−.12	.15	−.06	−.10
Syllable structure error ratio	**.78**	.18	−.08	.03	.09
Intonation error ratio	.01	**−.84**	.23	.15	.02
Articulation rate	−.10	**.79**	.12	.27	.07
Grammatical error ratio	.13	−.13	**.85**	.07	.21
Mean length of AS-unit	−.03	−.11	**−.56**	**.53**	−.11
MTLD	.14	.29	−.10	**.76**	.14
Word stress error ratio	−.11	−.22	.25	**.64**	−.14
Lexical error ratio	−.11	.06	.28	.01	**1.00**
Lambda	−.03	.37	.34	.00	**−.46**

Note. The factors were extracted with Promax rotation. Factor loadings above .40 are in bold.

Appendix 6

Correlation Matrix of Principal Components with Five-factor Solution Followed by Promax Rotation

Variable	1	2	3	4	5
1. PC1	—				
2. PC2	−.08	—			
3. PC3	−.09	−.04	—		
4. PC4	−.26	.03	–.04	—	
5. PC5	.34*	−.08	−.24	−.10	—

Note. * $p < .05$. Spearman's rank order correlation. PC stands for principal component.

Appendix 7

Results from a Principal Component Analysis of the 11 Objective Linguistic Features with 5-factor Solution Followed by Varimax Rotation

Variable	Factor loading				
	1	2	3	4	5
Vowel reduction error ratio	**.96**	−.10	−.01	.08	.02
Phonemic substitution ratio	**.92**	−.15	−.14	.16	−.01
Syllable structure error ratio	**.80**	.14	−.05	−.08	.16
Intonation error ratio	−.05	**−.83**	.17	.22	.06
Articulation rate	−.11	**.79**	.25	.09	−.05
MTLD	.06	.31	**.71**	−.16	.11
Word stress error ratio	−.26	−.17	**.66**	.21	−.20
Grammatical error ratio	.08	−.13	.05	**.81**	.08
Mean length of AS-unit	−.09	−.08	**.53**	**−.57**	−.02
Lexical error ratio	.06	.00	−.03	.20	**.91**
Lambda	−.14	**.40**	.01	.37	**−.54**

Note. The factors were extracted with Varimax rotation. Factor loadings above .40 are in bold.

Appendix 8

Results from a Principal Component Analysis of the 11 Objective Linguistic Features with 4-factor Solution Followed by Promax Rotation (Pattern Matrix)

Variable	Factor loading			
	1	2	3	4
Vowel reduction error ratio	**.98**	.06	−.03	.07
Phonemic substitution ratio	**.92**	.08	−.02	−.10
Syllable structure error ratio	**.78**	−.22	−.10	.11
Intonation error ratio	.01	**.88**	−.29	.03
Articulation rate	−.03	**−.60**	**.57**	.12
Word stress error ratio	−.08	**.47**	.33	**.40**
Lambda	−.06	−.24	**.78**	−.26
Lexical error ratio	.05	.04	**−.45**	−.11
Mean length of AS-unit	−.06	.09	−.25	**.79**
MTLD	.20	−.08	.17	**.70**
Grammatical error ratio	.21	.39	.39	**−.43**

Note. The factors were extracted with Promax rotation. Factor loadings above .40 are in bold.

Appendix 9

Correlation Matrix of Principal Components with Four-factor Solution Followed by Promax Rotation

Variable	1	2	3	4
1. PC1	—			
2. PC2	−.04	—		
3. PC3	−.26	.15	—	
4. PC4	−.27	−.08	.24	—

Note. * $p < .05$. Spearman's rank order correlation. PC stands for principal component.

Appendix 10

Results from a Principal Component Analysis of the 11 Objective Linguistic Features with 4-factor Solution Followed by Varimax Rotation

Variable	Factor loading			
	1	2	3	4
Vowel reduction error ratio	**.96**	.08	−.03	−.09
Phonemic substitution ratio	**.92**	.11	−.19	−.10
Syllable structure error ratio	**.79**	−.21	.04	−.16
Intonation error ratio	−.04	**.85**	−.03	−.21
Articulation rate	−.07	**−.55**	.17	.54
Word stress error ratio	−.21	**.47**	.37	**.43**
Mean length of AS-unit	−.14	.03	**.76**	−.10
MTLD	.09	−.09	**.68**	.26
Grammatical error ratio	.18	**.44**	**−.46**	.33
Lambda	−.10	−.16	−.22	**.72**
Lexical error ratio	.11	.01	−.13	**−.47**

Note. The factors were extracted with Promax rotation. Factor loadings above .40 are in bold.

Appendix 11

Results from a Principal Component Analysis of the 11 Objective Linguistic Features with 3-factor Solution Followed by Promax Rotation (Pattern Matrix)

Variable	Factor loading		
	1	2	3
Vowel reduction error ratio	**.92**	−.03	−.03
Phonemic substitution ratio	**.88**	−.06	−.18
Syllable structure error ratio	**.84**	.13	.11
Word stress error ratio	**−.42**	−.12	.13
Intonation error ratio	−.18	**−.80**	−.11
Articulation rate	−.06	**.76**	.09
Lambda	−.21	**.58**	**−.40**
Mean length of AS-unit	−.14	−.13	**.75**
Grammatical error ratio	.01	−.10	**−.62**
MTLD	.03	.22	**.57**
Lexical error ratio	.22	−.30	.02

Note. The factors were extracted with Promax rotation. Factor loadings above .40 are in bold.

Appendix 12

Correlation Matrix of Principal Components with Three-factor Solution Followed by Promax Rotation

Variable	1	2	3
1. PC1	—		
2. PC2	−.17	—	
3. PC3	−.16	.11	—

Note. * $p < .05$. Spearman's rank order correlation. PC stands for principal component.

Appendix 13

Results from a Principal Component Analysis of the 11 Objective Linguistic Features with Three-factor Solution Followed by Varimax Rotation

Variable	Factor loading		
	1	2	3
Vowel reduction error ratio	**.91**	.18	−.05
Phonemic substitution ratio	**.88**	.19	−.20
Syllable structure error ratio	**.83**	.01	.12
Word stress error ratio	**−.43**	.06	.12
Intonation error ratio	−.20	**.76**	−.22
Articulation rate	−.04	**−.76**	.19
Lambda	−.16	**−.62**	−.31
Mean length of AS-unit	−.19	.13	**.73**
Grammatical error ratio	.05	.08	**−.63**
MTLD	.00	−.20	**.60**
Lexical error ratio	.20	.33	−.02

Note. The factors were extracted with Varimax rotation. Factor loadings above .40 are in bold.

Appendix 14

Results from a Principal Component Analysis of the 11 Objective Linguistic Features with Two-factor Solution Followed by Promax Rotation (Pattern Matrix)

Variable	Factor loading	
	1	2
Vowel reduction error ratio	**.93**	.03
Phonemic substitution ratio	**.91**	−.07
Syllable structure error ratio	**.84**	.24
Word stress error ratio	**−.45**	−.09
Intonation error ratio	−.20	**−.81**
Articulation rate	−.04	**.75**
MTLD	−.03	**.46**
Lexical error ratio	.20	−.25
Lambda	−.14	.36
Grammatical error ratio	.08	−.36
Mean length of AS-unit	−.24	.19

Note. The factors were extracted with Promax rotation. Factor loadings above .40 are in bold.

Appendix 15

Correlation Matrix of Principal Components With Two-factor Solution Followed by Promax Rotation

Variable	1	2
1. PC1	—	
2. PC2	−.28	—

Note. * $p < .05$. Spearman's rank order correlation. PC stands for principal component.

Appendix 16

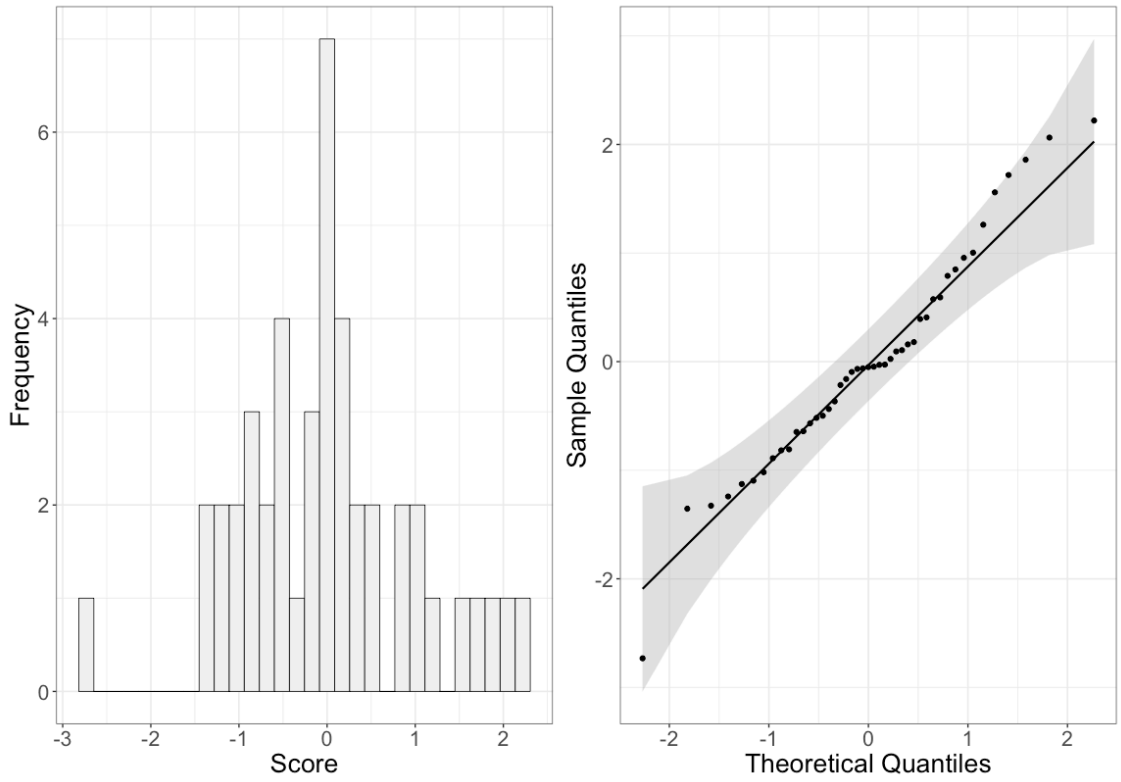
Results from a Principal Component Analysis of the 11 Objective Linguistic Features with Two-factor Solution Followed by Varimax Rotation

Variable	Factor loading	
	1	2
Vowel reduction error ratio	**.91**	−.17
Phonemic substitution ratio	**.89**	−.26
Syllable structure error ratio	**.82**	.06
Word stress error ratio	−.44	.00
Intonation error ratio	−.20	**−.77**
Articulation rate	−.04	**.76**
MTLD	−.03	**.46**
Lexical error ratio	.20	−.29
Lambda	−.14	.39
Grammatical error ratio	.08	−.38
Mean length of AS-unit	−.23	.24

Note. The factors were extracted with Varimax rotation. Factor loadings above .40 are in bold.

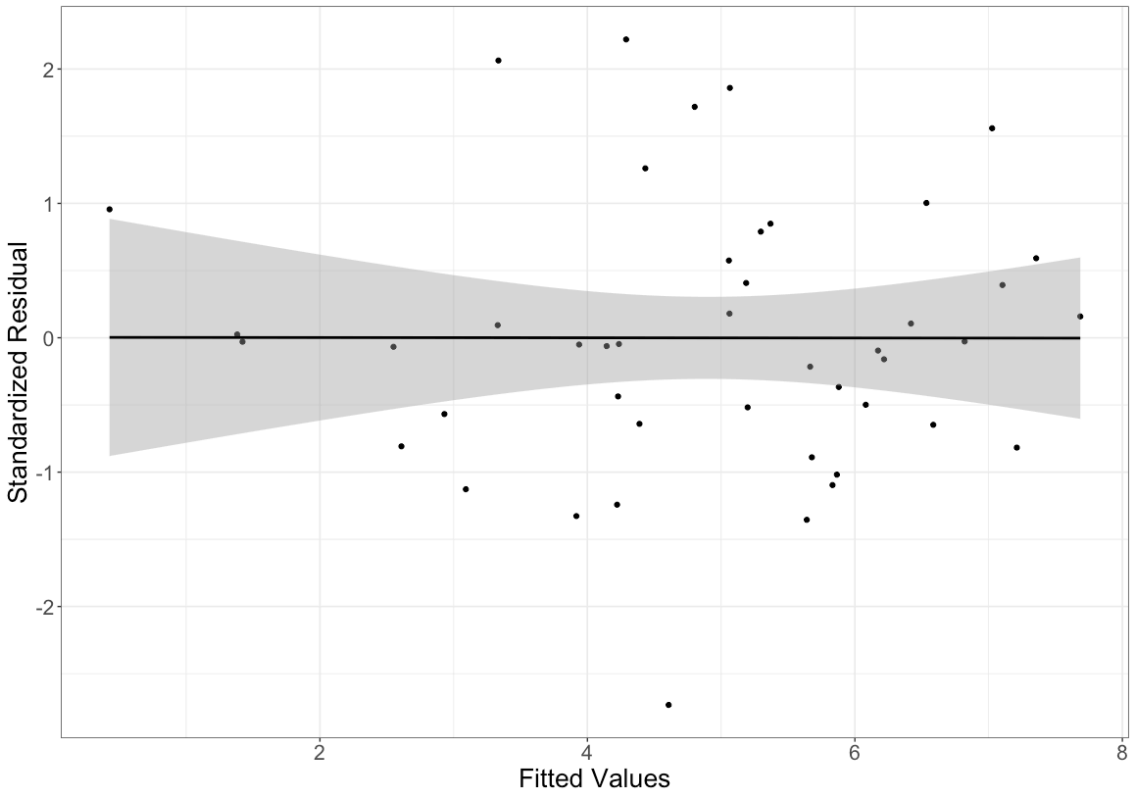
Appendix 17

Histogram and QQ-plot for Residuals of the Regression Model with Pronunciation and Lexicogrammar Extracted via Principal Component Analysis



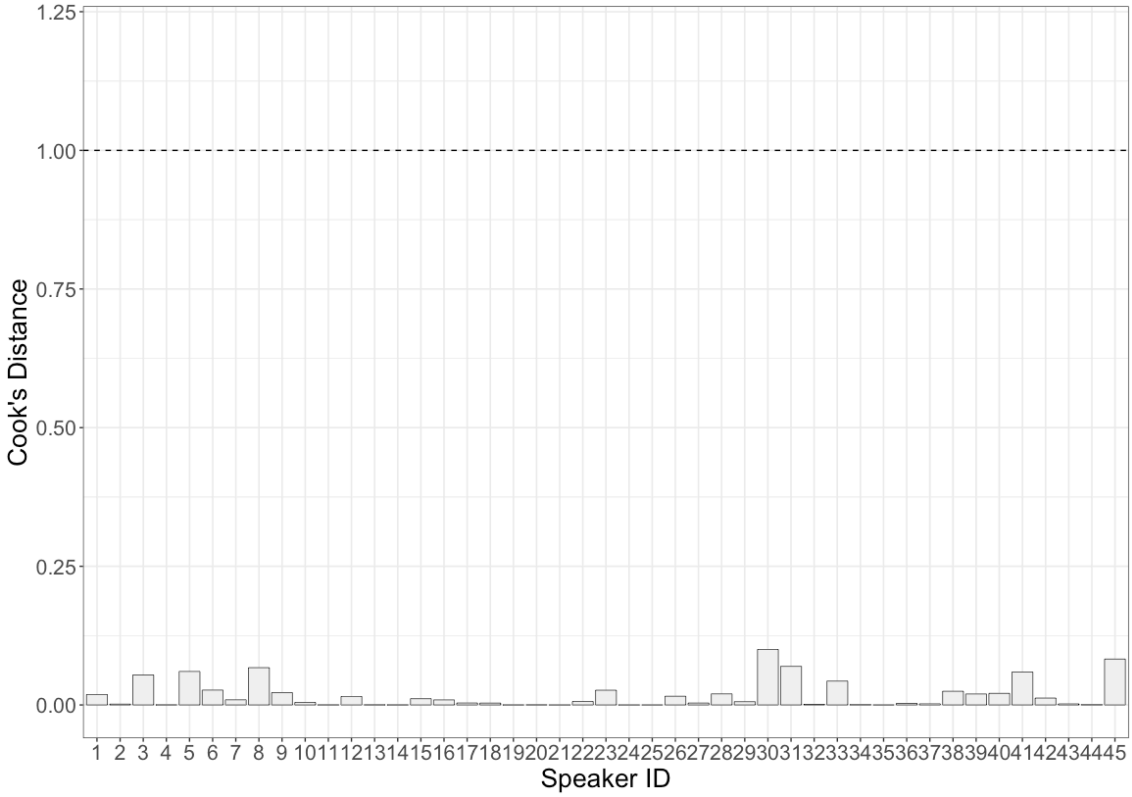
Appendix 18

Scatter Plot for Standardized Residual against Fitted Value for the Regression Model with Pronunciation and Lexicogrammar Extracted via Principal Component Analysis



Appendix 19

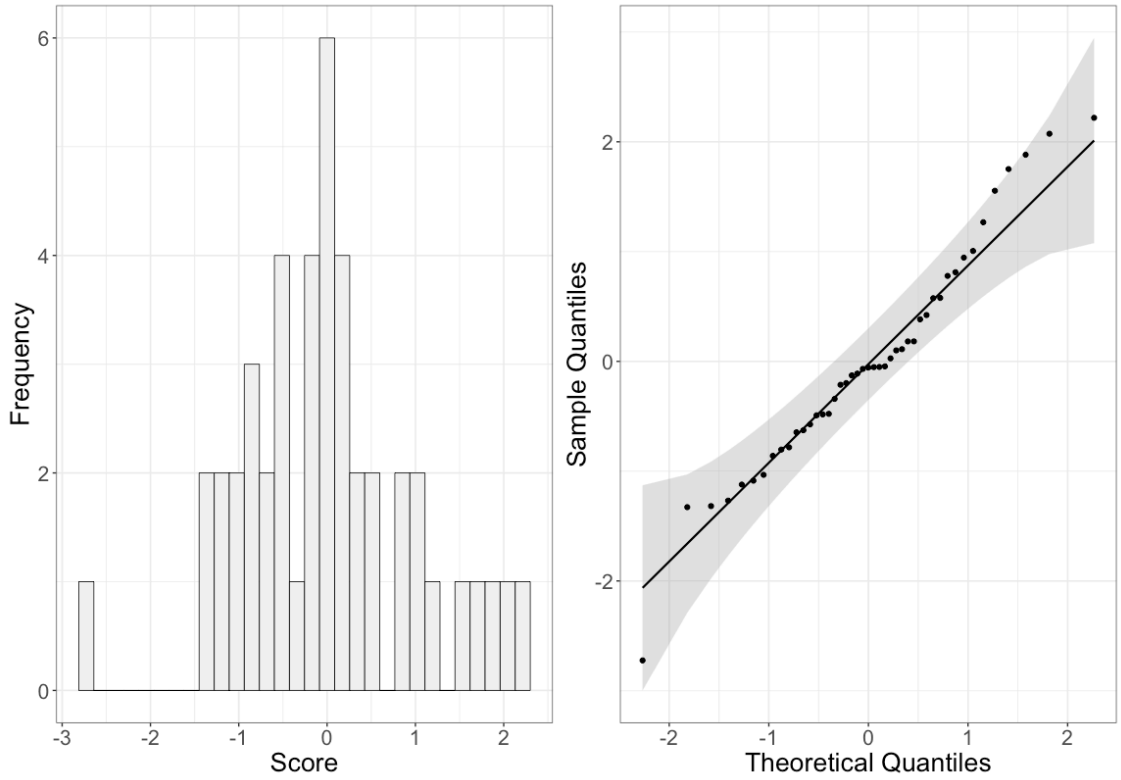
Cook's Distance for Each Observation in the Regression Model with Pronunciation and Lexicogrammar Extracted via Principal Component Analysis



Note. The dashed lines indicate reference points corresponding to Cook's distance of 1.

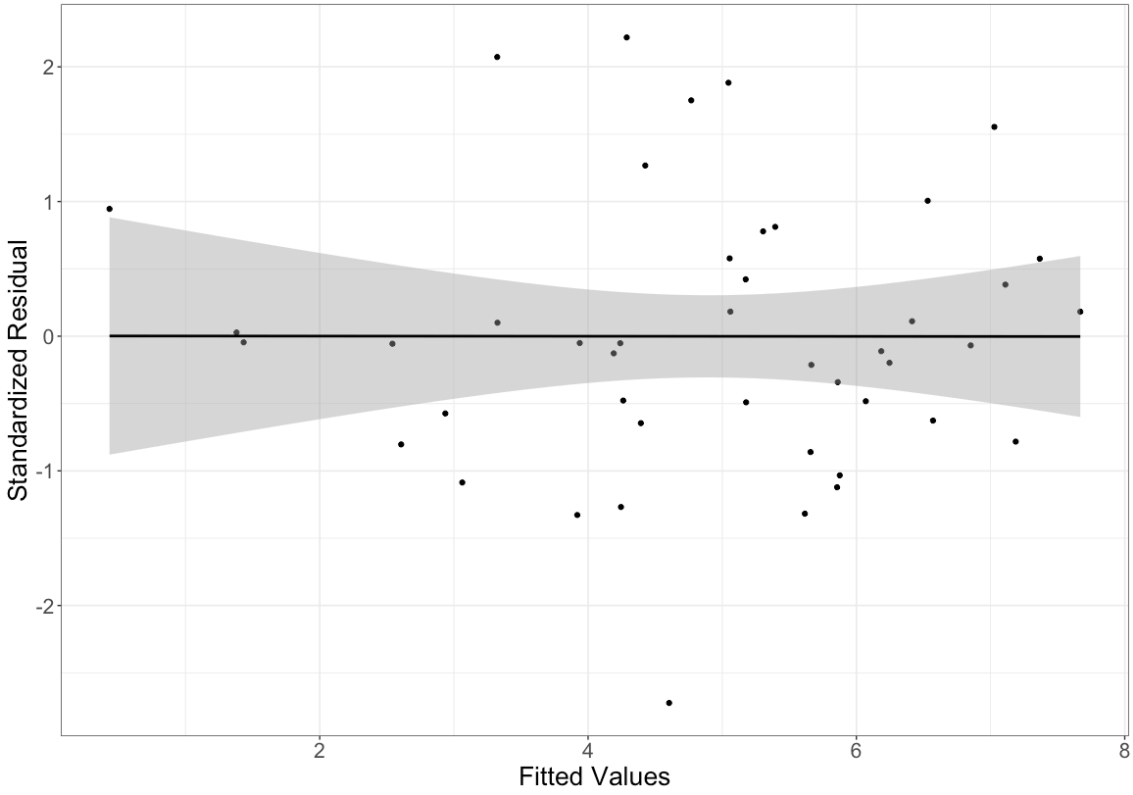
Appendix 20

Histogram and QQ-plot for Residuals of the Regression Model with Pronunciation and Lexicogrammar Extracted via Z-score Transformation Approach by Stanovich and West (1989)



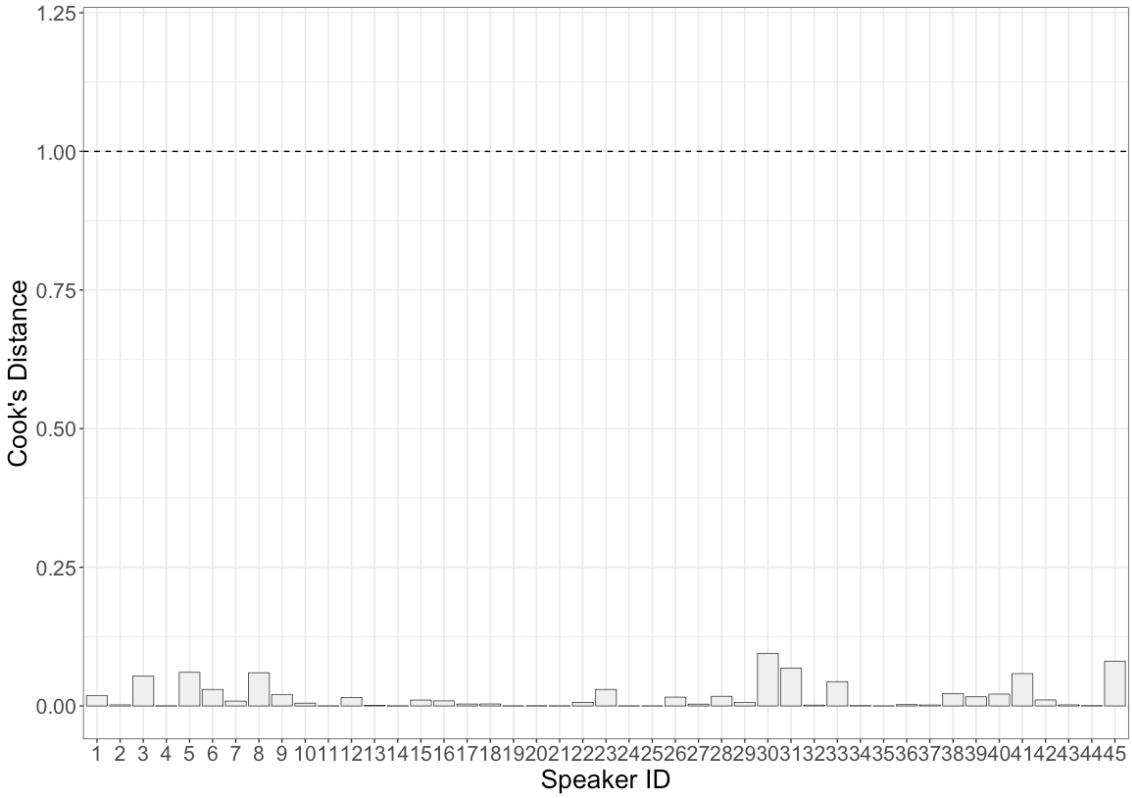
Appendix 21

Scatter Plot for Standardized Residual Against Fitted Value for the Regression Model with Pronunciation and Lexicogrammar Extracted via Z-score Transformation Approach by Stanovich and West (1989)



Appendix 22

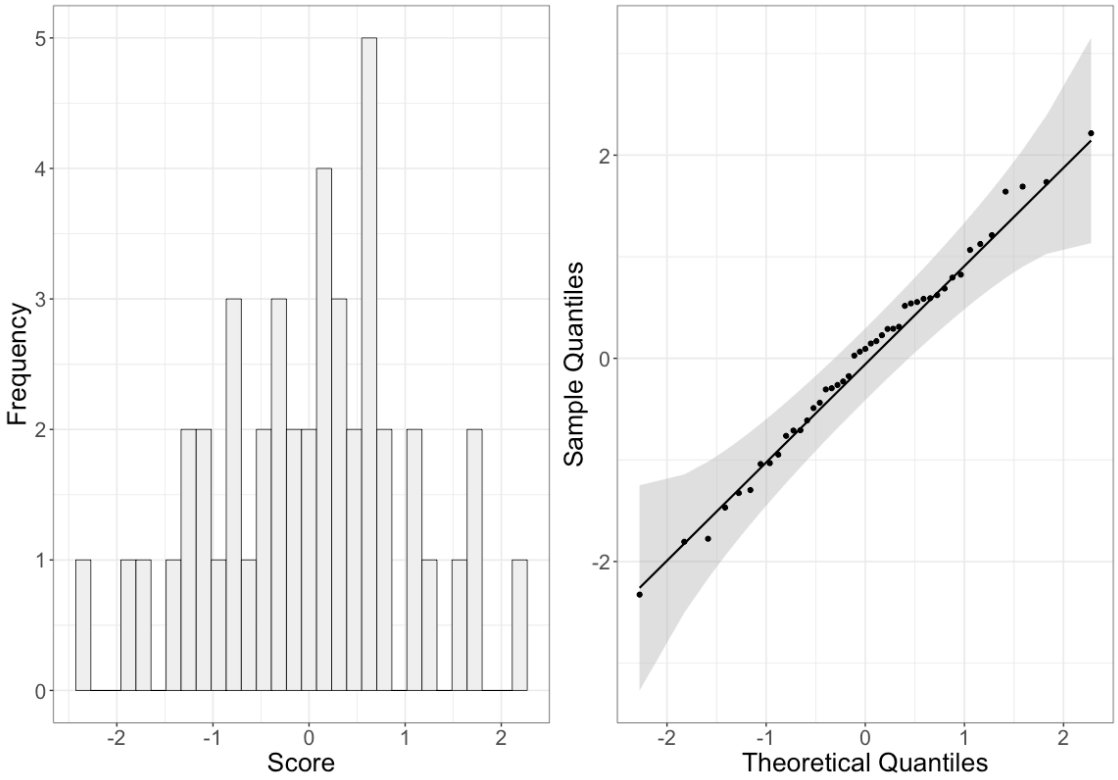
Cook's Distance for Each Observation in the Regression Model with Pronunciation and Lexicogrammar Extracted via Z-score Transformation Approach by Stanovich and West (1989)



Note. The dashed lines indicate reference points corresponding to Cook's distance of 1.

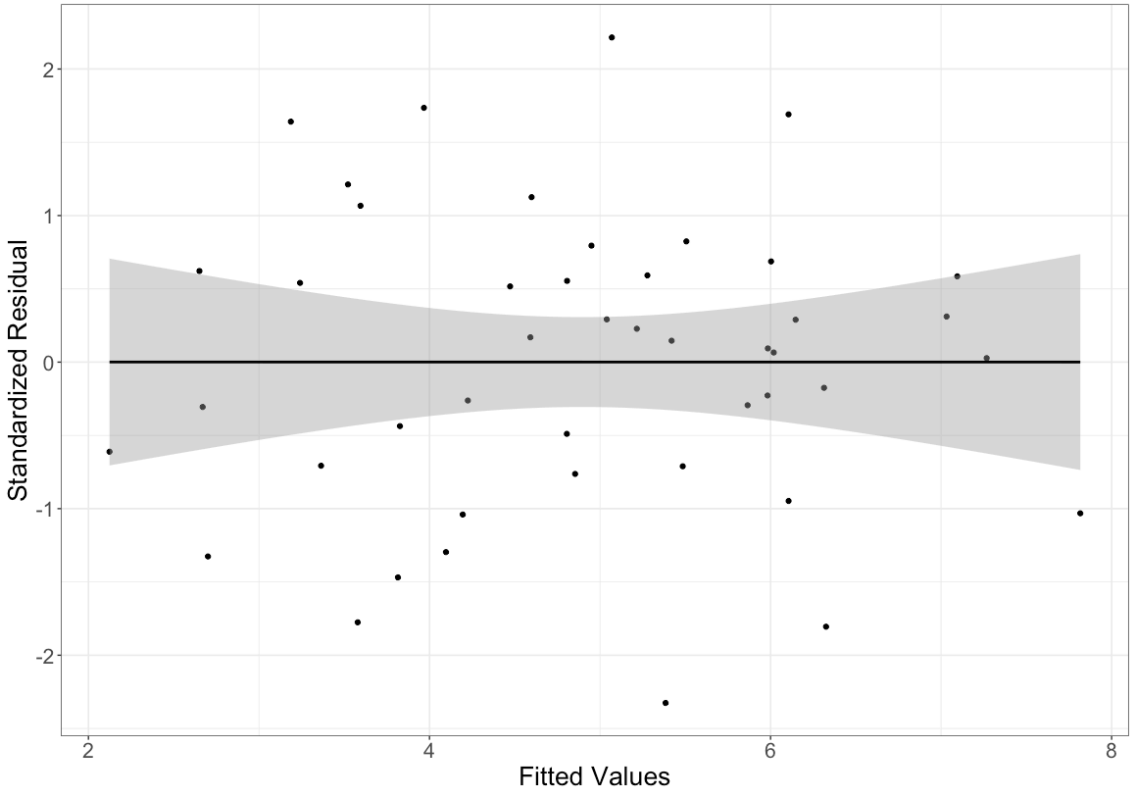
Appendix 23

Histogram and QQ-plot for Residuals of the Regression Model with Pronunciation and Lexicogrammar Extracted via Z-score Transformation Approach by Stanovich and West (1989)



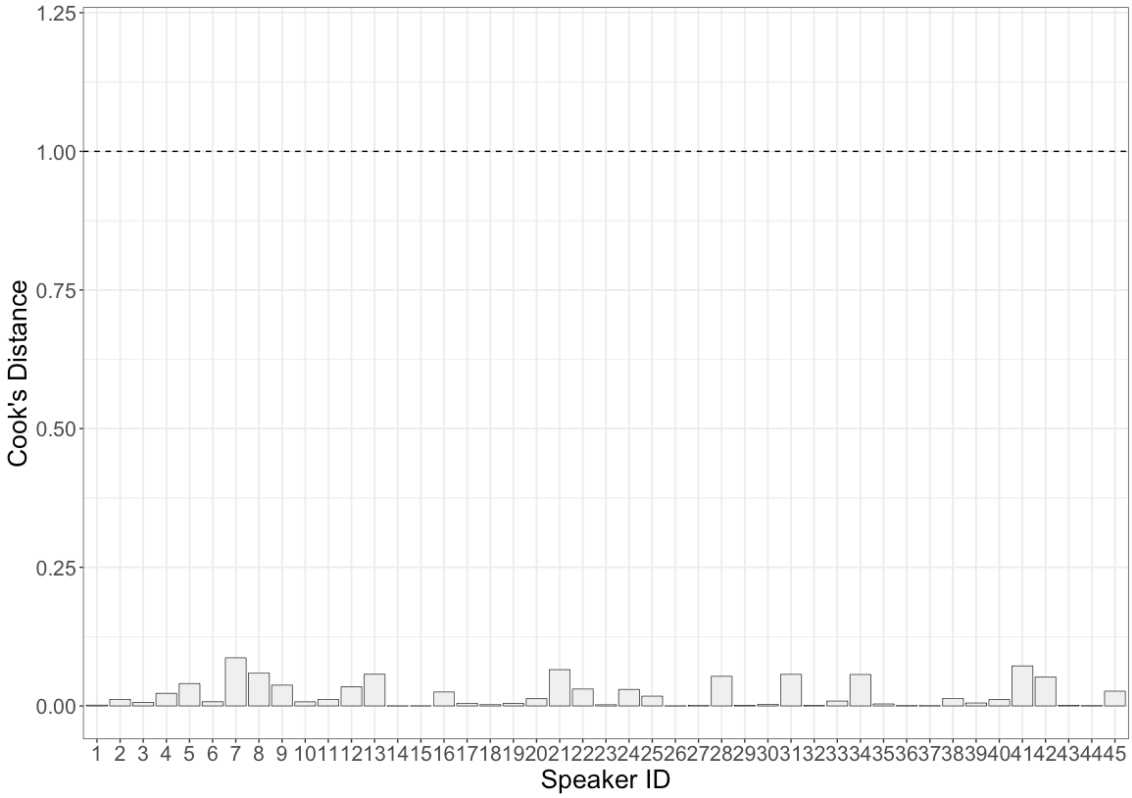
Appendix 24

Scatter Plot for Standardized Residual Against Fitted Value for the Regression Model with Pronunciation and Lexicogrammar Extracted via Z-score Transformation Approach by Stanovich and West (1989)



Appendix 25

Cook's Distance for Each Observation in the Regression Model with Pronunciation and Lexicogrammar Extracted via Z-score Transformation Approach by Stanovich and West (1989)



Note. The dashed lines indicate reference points corresponding to Cook's distance of 1.

Appendix 26

Results from Hierarchical Multiple Regression Analysis for Subjective Linguistic Features, with Pronunciation as the Initial Predictor, Followed by Lexicogrammar Extracted via Z-score Transformation Approach by Stanovich and West (1989)

Variable	*B*	95% CI for *B*		*SE B*	β	R^2	ΔR^2
		LL	*UL*				
Step 1						.763	.763***
Constant	4.89***	4.62	5.16	0.13	—		
Pronunciation	1.67***	1.39	1.95	0.14	.87***		
Step 2						.817	.054***
Constant	4.89***	4.66	5.13	0.11	—		
Pronunciation	1.15***	0.77	1.53	0.18	0.60***		
Lexicogrammar	0.69***	0.31	1.06	0.18	0.36***		

Note. CI = confidence interval; *LL* = Lower Limit; *UL* = Upper Limit.

*$p < .05$. ** $p < .01$. ***$p < .001$.

Appendix 27

Results from Hierarchical Multiple Regression Analysis for Subjective Linguistic Features, with Lexicogrammar as the Initial Predictor, Followed by Pronunciation Extracted via Z-score Transformation Approach by Stanovich and West (1989)

Variable	*B*	95% CI for *B*		*SE B*	β	R^2	ΔR^2
		LL	*UL*				
Step 1						.660	.660***
Constant	4.89***	4.57	5.21	0.15	—		
Lexicogrammar	1.55***	1.22	1.89	0.16	.81***		
Step 2						.817	.157***
Constant	4.89***	4.66	5.13	0.11	—		
Lexicogrammar	0.69***	0.31	1.06	0.18	.36***		
Pronunciation	1.15***	0.77	1.53	0.18	.60***		

Note. CI = confidence interval; *LL* = Lower Limit; *UL* = Upper Limit.

*$p < .05$. ** $p < .01$. ***$p < .001$.