v. 33, n. 1, p. 196–219, 2025
ISSN: 2237-2083 · DOI: 10.17851/2237-2083.33.1.196–219
Submetido em: 08/05/2024 · Aprovado em: 09/02/2025

REVISTA DE ESTUDOS DA LINGUAGEM

Pronunciation Assessment: Relating the CEFR's Phonological Control Scale to Intelligibility and Comprehensibility Measures

Avaliação de pronúncia: relacionando a Escala de Controle Fonológico do CEFR com Medidas de Inteligibilidade e Compreensibilidade

Thaisy da Silva Martins
Universidade Federal de Santa Catarina (UFSC) | Florianópolis | SC | BR
thaisy.sm@gmail.com
https://orcid.org/0000-0002-5638-9102

Rosane Silveira
Universidade Federal de Santa Catarina (UFSC) | Florianópolis | SC | BR
CNPq
rosanesilveira@hotmail.com
https://orcid.org/0000-0003-0329-0376

Abstract: Research in pronunciation learning and teaching over the past years has advanced the idea that working with the intelligibility and comprehensibility constructs is fundamental to developing pronunciation for successful communication. This study investigates the relationship between the Common European Framework of Reference for Languages - CEFR phonological scale (2018 version), and two measures commonly used in second language (L2) speech research, namely, intelligibility and comprehensibility rates. Speech samples from 16 Brazilian speakers of English were collected. Their speech samples were assessed by 14 listeners, teachers of English, in terms of intelligibility, comprehensibility and phonological control. The raters transcribed the speech samples, which generated an intelligibility score (number of words correctly transcribed); for comprehensibility, the raters assigned a level of difficulty in understanding the speech through a nine-point scale; finally, they used the CEFR scale to rate the level of phonological control for each speaker. Results show a highly significant correlation between intelligibility and comprehensibility, but not significant correlation between these two speech dimension variables and phonological control.

Keywords: speech assessment; intelligibility; comprehensibility; phonological control.

Resumo: A pesquisa sobre aprendizagem e ensino da pronúncia tem defendido, nos últimos anos, que sejam enfatizadas duas dimensões da fala: inteligibilidade e compreensibilidade. Essas dimensões são essenciais para o ensino da pronúncia que visa auxiliar no desenvolvimento de uma comunicação bem-sucedida. O presente estudo investiga a relação entre a escala fonológica do Quadro Europeu Comum de Referência para as Línguas - CEFR (versão 2018) e duas medidas comumente utilizadas na pesquisa sobre o desenvolvimento da fala em L2, a saber, inteligibilidade e compreensibilidade. Foram coletadas amostras de fala de 16 brasileiros falantes de inglês. Suas amostras de fala foram avaliadas por 14 ouvintes, professores de língua inglesa, quanto à inteligibilidade, compreensibilidade e controle fonológico. Os avaliadores transcreveram as amostras de fala, o que gerou um escore de inteligibilidade (número de palavras transcritas corretamente); para compreensibilidade, os avaliadores atribuíram um nível de dificuldade de compreensão da fala por meio de uma escala de 9 pontos; por fim, utilizaram a escala CEFR para avaliar o nível de controle fonológico de cada falante. Os resultados mostram uma correlação altamente significativa entre inteligibilidade e compreensibilidade, mas não foram obtidas correlações significativas entre as duas variáveis de dimensão da fala e controle fonológico.

Palavras-chave: Avaliação da fala; inteligibilidade; compreensibilidade; controle fonológico.

1 Introduction

In the field of oral proficiency assessment, pronunciation knowledge has often been assessed with a focus on accuracy or in a vague manner (Harding, 2017), rather than on whether L2 speakers can sustain a conversation with their interlocutors and establish communication without demanding too much effort from listeners. Responding to criticism against the 2001 Phonological Control Scale, the Common European Framework of Reference for Languages (CEFR) proposed an updated version of its phonological scale (2018), which seems to be aligned with the idea of focusing on successful communication.

Assessment is a relevant part in the process of learning a second language (L2). It contributes to the development of a learner's skills, and it helps teachers in the construction of knowledge with their students. In the specific domain of pronunciation assessment, Kang and Kermad (2018) demonstrate that, in the past, the assessment of a learner's oral performance used to focus on the accuracy of segmentals, having an idealized native speaker as the model. With time and understanding that native-like pronunciation was not an attainable and reasonable goal, pronunciation assessment started to take different paths, especially with the nativeness versus intelligibility principle proposed by Levis (2005).

According to Levis (2005; 2020), there are two principles in research concerning pronunciation: nativeness principle and intelligibility principle. The nativeness principle demonstrates the desire to sound native-like, as well as the prospect of achieving this goal, while the intelligibility principle defends the possibility and successfulness of communication in spite of one's accent. Intelligibility, in its turn, is defined by Derwing and Munro (2005, p. 385) as "the extent to which a listener actually understands an utterance". An important assumption is that having a strong accent does not mean that learners will not be understood by their interlocutor; the intelligibility concept holds that, if communication is established, there is no need for learners to sound native-like.

In addition to the intelligibility construct, research in the field of pronunciation assessment often investigates comprehensibility and accentedness as complementary dimensions. According to Derwing and Munro (2015), comprehensibility refers to how much effort the listener must make to understand what the speaker is saying. Accentedness also results from the listener's perception, and it relates to how the listener perceives certain pronunciation patterns as being different from those spoken by a specific speech community.

Thus, having intelligibility and comprehensibility assessment in mind, an important tool to assist this process is the Common European Framework of Reference for Languages (CEFR). This framework aims at providing guidelines to support language learning, teaching and assessment, according to the Companion Volume provided by the Council of Europe (2018).

Among other materials, the document features descriptor scales that can be helpful when assessing L2 learners' proficiency. The CEFR was updated to include a revised scale for phonological control, used for the assessment of L2 learners' oral proficiency. Claiming that the previous scale reinforced the view that accuracy and accent - and, therefore, the nativeness principle - were central to the development of L2 pronunciation, the phonological control scale was redeveloped to embrace the concept of intelligibility (Council of Europe, 2018).

Considering the CEFR scale for phonological control (2018 version), the concepts of intelligibility, comprehensibility, and the context of Brazilian learners of English, the objective of this study is to examine the relationship between the updated CEFR phonological scale, and two measures commonly used in L2 speech research, namely, intelligibility scores (measured as orthographic transcription of speech samples) and comprehensibility rates. One central question guides the study: How do raters' judgements of L2 learners' intelligibility and comprehensibility relate to the new CEFR scale for phonological control? We hypothesize that there is a correlation between intelligibility measures, comprehensibility ratings and the CEFR phonological control scale.

Having in mind that the updated CEFR scale for phonological control was released in 2018, there has not been much research on the subject (Khabbazbashi; Galaczi, 2020;Topal, 2019). Seeing that this is a valuable resource in the assessment of L2 learners, as well as a high -stake document used in language assessment, it is important to examine the adequacy of the update scale and relate it to the context of Brazilian learners of English. As potential study -abroad and work-abroad candidates, Brazilian undergraduate and graduate students often take standardized tests that evaluate English proficiency. Since the CEFR scales are used as a reference to assess proficiency in many high-stake proficiency tests, it is relevant to examine this recently revised resource for assessment implementation.

Having introduced the research context and stated our objective, the next section reviews relevant literature in the field of L2 speech, with a focus on the intelligibility and comprehensibility constructs, the CEFR guidelines, and selected empirical studies that address pronunciation assessment. Then, we explain the method employed to collect and analyze speakers' and listeners' data. Finally, we present and discuss the results of the correlational analysis. The article ends with tentative conclusions and directions for further studies and classroom implications.

2 Review of literature

In this section we present an overview of the studies connected with pronunciation assessment, including important constructs such as intelligibility and comprehensibility, accent, pronunciation assessment, raters, and rating scales. After that, we describe the CEFR phonological control scale and empirical studies examining the CEFR impact.

2.1 L2 pronunciation assessment dimensions

Following Derwing and Munro (2015, p. 2), we can define pronunciation as "the ways in which speakers use their articulatory apparatus to create speech". Pronunciation is an essential component of L2 oral development, and it encompasses knowledge about segments (vowels and consonants), suprasegments (e.g., stress, intonation, connected speech phenomena), as well as voice quality features (e.g., tone). Successful L2 communication involves active collaboration between speakers and listeners in order to accommodate differences in pronunciation.

Derwing and Munro (2015) identify L2 speech dimensions that affect successful communication and that should be considered when assessing the oral component of L2 proficiency, namely, intelligibility, comprehensibility, accentedness, and fluency. The authors define intelligibility as the extent to which a listener understands the speaker's intended message. On the other hand, comprehensibility is related to the degree of effort required from the listener to understand what the speaker is saying. Accentedness also results from the listener's perception, and it relates to how certain pronunciation patterns are perceived as being different from those spoken by a specific speech community. Finally, fluency is a speech dimension related to speech rate (speed) and fluidity, encompassing the frequency of occurrence of pauses and hesitation markers.

A number of studies have investigated the intelligibility construct in the field of pronunciation assessment, often in connection with other relevant constructs such as comprehensibility, accentedness, linguistic features, and rater and speaker characteristics.

Derwing and Munro (1995b) analyzed accentedness, comprehensibility and intelligibility in the speech of second language learners. Native speakers of English judged these speech dimensions assessing speech produced by non-native speakers and the results suggested that, even when speakers were judged as having a strong foreign accent, their intelligibility was not affected. These results corroborate the claim advanced by Levis (2005, 2020) that communication can be intelligible even when the speaker has a strong non-native accent. Furthermore, Derwing and Munro's (1995a, 1995b) study have been highly influential in the field of L2 speech research, as they have consistently shown that while intelligibility, comprehensibility, and accentedness are related constructs, they are also independent from each other.

Bent and Bradlow (2003) investigated the influence of the native language background on intelligibility assessment. Their results demonstrated that native listeners considered native speakers more intelligible than non-native speakers; moreover, a "matched interlanguage speech intelligibility benefit" could be seen in the results. This benefit means that non-native speech is more intelligible to the non-native listener because both listeners and speakers share the same native language, and this shared knowledge impacts intelligibility (Bent; Bradlow, 2003).

A study conducted by Kang *et al.* (2017) aimed at examining the relationship between phonological features of the L2 speech (segmentals and suprasegmentals) and different measures of intelligibility, as well as the correlation between these measures and listeners' comprehension scores in the TOEFL exam. The participants were high-proficient English users. The listeners assessed the speakers' intelligibility through five different measures: true/false statements, scalar ratings, perception of nonsense sentences, perception of filtered sentences and orthographic transcription. The authors concluded that comprehensibility of nonsense sentences was effective in measuring the intelligibility of speakers from different backgrounds because they "require a phonemic level of speech processing and, to some extent, knowledge of sound cooccurrences" (Kang *et al.*, 2017, p. 138).

Silveira and Silva (2018) investigated the intelligibility of English word-final codas (e.g., 'bed' pronounced as [bɛdʒ]) produced by Brazilian learners and assessed by listeners from different L1 backgrounds. The study correlated intelligibility with the listeners' second language (L2) proficiency level, familiarity with speakers' L1, and length of residence in the speakers' country. The results indicate that certain types of coda modification have a negative effect on intelligibility, and that semantic information present in the carrier phrases improves intelligibility in some cases. Furthermore, listeners' familiarity with the English spoken by Brazilians helps them to perform better on the intelligibility task, which demonstrated that accent familiarity contributes with successful communication between speakers from different L1 backgrounds, as accent familiarity brings awareness about L2 pronunciation patterns that might hinder intelligibility.

Comprehensibility is another important concept in the assessment of pronunciation, and it is often explored alongside intelligibility and accentedness. Comprehensibility is usually measured through Likert-scales. Derwing and Munro (1995a) investigated the relationship between speech processing time (measured in terms of how long it took listeners to respond), accentedness and comprehensibility rates, provided by English native speakers. The results showed a relationship between comprehensibility and response time, suggesting that native speakers take processing time into account when evaluating the speech of non-native speakers.

Trofimovich and Isaacs (2012) investigated what linguistic features (phonology, fluency, lexis/grammar, and discourse) are related to accent and what features are related to comprehensibility. Speech samples were provided by French native speakers and evaluated by inexperienced raters and experienced teachers of English. The results demonstrate that phonological aspects, such as segmental accuracy, are more related to accentedness, while grammatical and lexical errors are linked to comprehensibility. This suggests that segmental errors affect accentedness ratings, while grammatical and lexical errors affect comprehensibility ratings.

Foote and Trofimovich (2018) investigated the role of the listeners' native language in the process of assigning comprehensibility ratings. The listeners were L2 English speakers from Mandarin, French, Hindi, and English backgrounds. They rated speech samples of L2 English speakers from Mandarin, French and Hindi backgrounds. The results demonstrated that the speakers' L1 background must be considered as a factor when evaluating comprehensibility. Besides, when the listeners shared the same L1 with the speakers, they tended to make positive comments about their speech; when they did not share the same L1 background, they tended to make negative comments, suggesting a relation between comprehensibility and L1 background.

Saito, Trofimovich and Isaacs (2015) examined the correlation between comprehensibility and accentedness for learners from different proficiency levels. Japanese speakers of beginner, intermediate and advanced levels of English completed a speech elicitation task that involved describing an image. Their speech was assessed by inexperienced native speakers of English in the domains of comprehensibility and accentedness. Experienced native -speaker raters evaluated the speech samples focusing on linguistic analyses of phonological, lexical, and grammatical characteristics of speech. The results showed that comprehensibility was related to segmental, prosodic, temporal, lexical and grammatical aspects of L2 speech, while accentedness was related mainly to segmental accuracy. This study also contributes to the accepted view in the area that a speaker with high phonological, lexical, and grammatical proficiency can be comprehensible while still having an accented speech. Regarding the differences on comprehensibility assessment across proficiency levels, the authors highlight that for beginner to intermediate learners, prosody, temporal variables, and lexical accuracy are the main targets; for intermediate to advanced learners, the listeners tend to focus on segments, prosody, and grammatical accuracy.

Having a strong L2 accent may result in miscommunication, due to the speaker's pronunciation of segmental aspects (vowels and consonants) or suprasegmental aspects (e.g., stress or intonation), and it might create confusion, irritation, or even prejudice against the speaker (Derwing; Munro, 1995a). Derwing and Munro (1997) examined the relationship between accentedness, comprehensibility and intelligibility, concluding that accentedness ratings are harsher than comprehensibility ratings, which in turn are harsher than intelligibility scores.

2.2 Pronunciation assessment: measurement and rater issues

Pronunciation assessment can vary in method and type of measurement. Kang and Kermad (2018) explain that pronunciation can be assessed either by human beings or by machines. While machine assessment relies on acoustic parameters, assessment provided by listeners/raters can use methods such as rating scales, orthographic transcription, true/false questions, cloze tests with nonsense sentences, or comprehension questions. Human raters can be biased due to a variety of reasons, which can impact the process of evaluation and thus must be considered (Kang; Kermad, 2018).

According to Derwing and Munro (2015), in the context of standardized tests - in which there is the use of scales to assess pronunciation - there is also a disparity in the results, due to the differences that permeate human evaluators; the authors highlight that this can be a serious issue due to the fact that standardized tests (such as the Test of English as a Foreign Language - TOEFL - and the International English Language Testing System - IELTS) are often used to make decisions about students' admission in school and work programs.

A study conducted by Kang et al. (2019) examined the effects of raters' background on the evaluation of English non-native speakers, and how a brief training could help neutralize this impact on oral assessment. The participants were naive raters with no formal experience in evaluating oral proficiency, and they all varied in language background. They holistically evaluated speech samples produced by TOEFL examinees. The results demonstrated that English native-speakers tended to be less strict than non-native speakers when assigning rates. Moreover, frequent contact with a certain accent impacted the raters in the sense of being more lenient towards familiar accents. The study also demonstrated that a training session - which happened online - helped to balance the assessment, that is, after the training the raters' interrater reliability increased.

Rating scales rely on a listener's perception and judgment and many factors can influence the outcome of this rating. Kuiken and Vedder (2014) highlight factors such as the speaker's proficiency level, type of task, topic of the assignment, rating experience, familiarity with rating scales, and training. Toffoli *et al.* (2016) add relevant factors such as raters' strictness, task difficulty, assessment criteria and scale descriptors. A rating scale is a sort of framework that enables a listener to judge specific language traits from a speech sample, while following a structure in order to minimize possible interferences such as seen above (Isaacs; Thomson, 2012).

Usually, researchers work with 9-point Likert scales to assess comprehensibility, where 1 refers to "no difficulties" and 9 to "extremely difficult". According to Isaacs and Thomson (2012), 9-point scales are often chosen due to their versatility, meaning that they can be used with learners from any L1 and by inexperienced raters who have no background in linguistics. The authors also highlight that Cronbach's alpha coefficients (a statistical test to assess interrater reliability) are often high when using this method, which is a good indication of the validity of 9-point scales as a method to assess speech production, considering that there must be an agreement between the raters in order for their ratings to be reliable. Moreover, the authors argue that without proper rater training or rigorous criterion-referenced standards, the possible interpretations generated by the use of the rating scales by the listeners can happen only within the same study, not across studies.

The debate between the performance of expert raters vs. naive raters has led to different conclusions. Isaacs (2013) discusses that having naive raters assess learners' speech may be interesting since these are the type of people that L2 learners will encounter in real-life and have a conversation with. The author demonstrates that recruiting experienced raters may be unnecessary because naive raters have demonstrated to assign reliable rates for L2 speech; however, the author stresses that there may be different outcomes depending on the objectives of the assessment, because experienced and novice raters may approach the rating task differently and thus produce different results. In the case of this study, it is more appropriate to use experienced raters, or L2 teachers, to assess the speakers' oral proficiency because this group of listeners are more familiar with rating scales and descriptors, such as the one provided by the CEFR, and our study does not provide listeners with training on how to use the rating scales.

2.3 The CEFR and its phonological control scale

The CEFR is an important guideline for language learning, teaching, and assessment (Harding, 2017). It is a source for the development of language syllabuses, curriculum guidelines, examinations, textbooks and so on across Europe, and it has been influential in other contexts as well (Figueras, 2012), including in Brazil. The framework also provides descriptors for levels of proficiency and for assessment of proficiency subcomponents. The CEFR embraces the concepts of action-oriented approach, communicative language competence, tasks and strategies for language learning. The framework is also based on the plurilingualism idea, in which a learner experiences languages considering its cultural aspects, constructing an interrelation and interaction among languages (Council of Europe, 2001). The CEFR guidelines had a huge impact on language learning and teaching, on the development of textbooks and high-stake proficiency tests worldwide; it contributed with instruction and information, for language teachers and professionals, on language proficiency, language teaching, learning and assessment, as well as in how to operationalize and apply these constructs (Quevedo-Carmargo, 2019).

The Common Reference Levels were developed to help with the description and measurement of levels of proficiency, creating a scheme to help in the comparison between different systems and standardized language tests. The reference levels provided by the CEFR are breakthrough and waystage for the basic user (A1 and A2), threshold and vantage for the independent user (B1 and B2) and effective operational proficiency and mastery for the proficient user (C1 and C2). The CEFR is composed of a global proficiency scale and illustrative scales. The illustrative descriptor scales comprise five linguistic competences: vocabulary range, grammatical accuracy, vocabulary control, phonological control, and orthographic control.

The focus of this study and the only descriptor scale that is going to be used in the present research is the phonological control scale. The description for phonological competence in the 2001 CEFR guidelines defines it as the production and perception of "the sound-units of the language, and their realisation in particular contexts; the phonetic features which distinguish phonemes; the phonetic composition of words; sentence stress and rhythm; intonation; vowel reduction; strong and weak forms; assimilation and elision" (Council of Europe, 2001).

In the 2001 version, the phonological control scale is concise and does not describe each level in detail. A1 level, for example, is described as "pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by native speakers used to dealing with speakers of his/her language group" while C2 and C1 levels are described as: "can vary intonation and place sentence stress correctly in order to express finer shades of meaning" Council of Europe, 2001, p. 117). These descriptions are quite vague and can lead to misplacement of candidates. The CEFR phonological scale was updated in 2018 in order to embrace more concepts related to oral proficiency and to provide more detailed guidelines for assessment. Thus, in the descriptors of the 2018 phonological scale, there can be seen the integration of the articulation, prosody, accentedness and intelligibility constructs. The updated descriptors in the Phonological Control scale, displayed in Annex 1, gives special attention to intelligibility, sound articulation and prosody, diminishing the focus on accuracy and accentedness (Council of Europe, 2018). An improvement can be seen in the updated version, as it is more in-line with current research and studies in the area of pronunciation teaching and assessment. A new CEFR companion volume was published in 2020 with some alteration in the scales; however, the phonological control scale remained the same as the 2018 version.

Figueras (2012) investigated the impact of the CEFR for language learning, teaching, and assessment. The author demonstrated that the CEFR was developed and published during a time when language professionals were trying to describe and establish guidelines to help to inform language learning, teaching and assessment, especially on how to categorize language learning from "lack of knowledge" to "effective mastery". One of the main contributions, according to the author, was the definitions for each level provided by the framework (A1-C2). The language descriptors were also rapidly adopted, especially to aid in the development of language learning programs, from different contexts and places. One important influence of the CEFR was in relation to how it describes learners' progress. Instead of stating what learners cannot do at a specific level, the CEFR descriptors highlight what the learners can do, the now famous can-do statements. Moreover, the author demonstrates that the CEFR can also be misused depending on how and where the descriptors are going to be implemented.

Deygers *et. al* (2018) investigated the impact of the CEFR on European university admissions. The authors conducted interviews with representatives of 30 organizations, professionally involved with language testing and in the development of language tests for university entrance. The interview focused on university entrance policy, entrance tests, and personal opinions about the CEFR and university entrance language tests. The results demonstrate that the CEFR has a great impact on university entrance in Europe, and that many of the respondents mentioned having a positive view towards the framework. However, the authors display some controversies regarding the use of the scales on entrance tests: sometimes this instrument is misused because "[...] in many contexts it now serves as a self-administered seal of quality. It can give university admission officers a semi objective tool to control university entrance, and it may allow test developers to claim a link to a certain level without having to offer any kind of proof for this." (Deygers *et. al*, 2018, p. 10). This article illustrates the importance and impact of the CEFR while also discussing its potential misuse.

This brief review demonstrates that the Common European Framework is to serve as a reference for language teaching, learning and assessment (Council of Europe, 2018), and it is important to adapt it to one's needs. Furthermore, users should consider the implications of using the assessment scales and what types of negative results they might bring, especially because the updated Phonological Control Scale including intelligibility, comprehensibility, and articulation constructs in its design is very recent. The present study explores the Phonological Control Scale potentials and limitations by correlating it with other consolidated measures of L2 speech, namely, intelligibility and comprehensibility.

3 Method

In this section, the participants, instruments, and materials of this study will be presented in detail, as well as the procedures for data collection and data analysis. This research was conducted through speaking tasks and listening tasks, with the objective of examining the relationship between the CEFR phonological scale (2018) and intelligibility scores (measured as orthographic transcription of speech samples) and comprehensibility rates.

3.1 Participants

The study gathered data from a group of speakers and a group of listeners. Speakers were invited to participate in this research through social media, contact with English teachers and schools from the city where the study took place (south of Brazil). Speakers were 16 Brazilian learners of English as an L2, undergraduate students from different majors. Their age ranged from 17 to 29 (m = 21.87, sd = 2.41), and they reported having experienced a different number of years learning English, ranging from 5 to 20 years (m = 11.06, sd = 3.71). Similarly, they reported a different number of hours using English every day: seven reported 2-6 hours, five reported 2 hours, two reported 6-10 hours, and two speakers reported 10 hours or more. Speakers' proficiency was estimated through the Oxford Placement Test, and the study has speakers with proficiency levels ranging from intermediate to advanced (B1 = 9; B2 = 5; C1 = 2), being most of them independent users of English according to the CEFR proficiency levels.

The study also gathered data from a group of listeners who assessed the speech samples provided by the group of speakers. The listeners of this study were 14 English as an L2 teachers, in order to have experienced raters with a background on language assessment. Listeners were also contacted through social media and personal connections with graduate students and in-service teachers from language institutes to be volunteers in this research. They answered a questionnaire to report their language use and teaching experience. Listeners' age ranged from 22 to 35 (m = 29.07, sd = 4.39. They reported having been teaching English from 3 – 20 years (m = 9.21, sd – 4.80) in a range of contexts (language institutes, private classes, regular schools, undergraduate programs, bilingual schools).

3.2 Instruments and materials

This section brings information about the consent forms, background questionnaires, proficiency test, speech elicitation task, and listening tasks. The research was submitted to the Ethics Board[1] and received the approval to start the data collection. There were two different consent forms for the participants of this research: one for the speakers and one for the listeners. The participants of this study, after reading and signing the consent form, answered a background questionnaire. For the speakers, there were questions related to their language learning (number of years) and also to their daily language use. For the listeners, the questions addressed their experience as English teachers and working contexts.

To estimate the speakers' proficiency before assessing their intelligibility, an adapted version of the paper and pen version of the Oxford Placement Test was administered online through a Google form. The selected version of the Oxford Placement Test assesses reading, vocabulary, and grammar, and follows the CEFR levels (A1-C2) (Allan, 2004). The adapted version included 60 questions. Further information about all instruments is available in Martins (2022).

Speech samples were elicited with an Image Description test completed by the speakers. This test (Image 1) consisted of an image of a working space with people interacting and working together, and the speakers were expected to describe the image as much as they could. Image description tests have been used in the literature as an efficient instrument to

[1] CAAE: 48418621.9.0000.0121.

elicit speech (Derwing *et al*., 2008; Isaacs; Trofimovich, 2012; Silveira; Silva, 2018; Silveira; Martins, 2020). Since the raters were expected to transcribe speech samples produced by the speakers, this type of free-speech sample is more suitable because it prevents listeners from guessing and getting used to the words, as they would if it was a reading aloud task, for example. The speech samples were collected online, and audio recorded using the Zoom platform.

Image 1 – Speech Elicitation Image

Source: Martins (2022)

The image was selected based on the elements available for the speakers to describe, that is, an image with enough possibilities for description without being too overwhelming. It was expected that the speakers would mention the people in the picture, the objects, the actions, the possible reasons for the situation, the colors, the background, etc. Low-proficient speakers could focus on the colors and forms present in the picture, while high-proficient speakers could explore the details and make possible abstract inferences such as conversation topics, possible relationships between the characters and so on.

As shown in Table 1, the raters listened to the speech samples and assessed them in terms of intelligibility (orthographic transcription of utterances), comprehensibility (9-point rating scale[2]), and phonological control (6-point rating scale with CEFR descriptors). The intelligibility score consisted of the percentage of words correctly transcribed by the listeners, and the comprehensibility rating consisted of listeners' perception of difficulty to understand the utterances. Finally, the listeners used the CEFR descriptors to rate phonological control, which is part of the speaking proficiency scales (CEFR, 2018). The three assessment tasks completed by the listeners were presented on Google form, organized in different sections. The first section was developed to gather listeners' background information. In the second section, the listeners assessed the speakers in terms of intelligibility, comprehensibility and they also assigned them a CEFR level for phonological control. For each speech sample, the form provided the listener with a space for orthographic transcription, a Likert scale for the comprehensibility measurement, and a chart with the CEFR descriptors for the raters to consult

[2] The choice of a 9-point scale is based on the literature related to intelligibility and comprehensibility (Thompson, 2018).

and, right after this chart, a multiple-choice item from which the listeners could select the appropriate level of the phonological scale. These steps were repeated for each speaker, in a total of 16 parts within the Google Form. Further information about the assessment form is available in Martins (2022).

Table 1 – Content of listeners' assessment form

Components	Content
Speaker audio file	1 audio file for each speaker, accompanied by image being described.
Intelligibility task	Box for orthographic transcription of the speech sample.
Comprehensibility rating scale task	9-point scale for rating difficulty to understand speech sample.
Phonological control scale task	Complete phonological control scale and list of 6 descriptors for assigning level of control for speaker.

Source: the authors

3.3 Procedures for data collection

All prospective participants received an email with the consent form and the background questionnaires. After reading and signing the consent form and completing the questionnaire, they received another email with instructions to schedule an online, individual meeting through Zoom, with one of the researchers, to complete the other data collection steps.

Data were first collected from speakers. During the individual Zoom meeting session, each speaker received a link to a Google form for the Oxford Placement Test, which they completed during the meeting, with their cameras turned on. In the same meeting, after finishing the placement test, they completed the speech elicitation task (image description). For this purpose, the researcher in charge of data collection shared a slide presentation with the instructions for the speech elicitation task. Each speaker had a maximum of 30 seconds to plan the description, without taking notes. Speakers were instructed to describe the image, and the complete sessions were recorded (video and audio files) using Zoom functions. The speakers were informed that their image was not to be used at any time, only the audio with their speech sample would be the object of analysis. The speakers needed around 60 minutes to complete all the tasks.

The speech samples were normalized using the Audacity software and prepared to be part of the listening tasks that allowed assessing the speech samples. The normalization process involved increasing the sound volume and removing background noises.

The second step involved asking listeners to assess the speech samples. Prospective listeners who volunteered to contribute with the study received an email with instructions to complete the listening tasks remotely, at their own pace. The email included a link to the Google Form where all listening tasks were organized, with 16 parts containing the speech samples of each speaker and the respective sections to assess intelligibility, comprehensibility, and phonological control using the CEFR scale. Listeners were asked to wear headphones when performing all the assessment tasks.

For the intelligibility assessment, the raters were asked to transcribe every word of the speech samples. Following Derwing and Munro (1997), listeners were oriented to transcribe everything exactly as they heard, without making any type of correction when transcribing. Their transcriptions were automatically saved in Google Forms. Listeners were able to listen to the speech samples as many times as they found necessary as they were performing the transcription.

For the comprehensibility measurement, there was a Likert scale for the raters to decide how difficult it was to understand the utterances. In the Likert scale ranging from 1 to 9, "1" referred to "extremely difficult" and "9" referred to "no difficulties"[3]. The listeners filled out the comprehensibility scale after finishing the transcription of each speech sample and they were allowed to listen to the speech sample again if they wanted to. The listeners were oriented to use the entire scale, as the scale represented a range of ability levels.

After completing the intelligibility and comprehensibility tasks, the raters were asked to assign phonological control levels to each speaker, using the CEFR descriptor scale (2018) for phonological control, which showed a range of phonological control levels, from A1 to C2. The CEFR phonological control scale was presented to the listeners using the same Google Form that was created for the intelligibility and comprehensibility tasks, but it appeared as the final task. The listeners were oriented to reflect about their choices while using the scale, as well as to pay close attention to the descriptors to make a conscious choice regarding the speakers' phonological control level.

3.4 Procedures for data analysis

To obtain a proficiency measure for each speaker, we accessed the Google Form spreadsheet containing their answers to the Oxford Placement Test. We examined the answers of each speaker with the answer-sheet that accompanies the test and assigned the proper proficiency level according to the instructions (Allan, 2004). All speakers received either a B1 level, B2 or C1, for their reading, vocabulary, and grammar L2 proficiency.

The intelligibility measure involved comparing the orthographic transcriptions provided by each listener with the speech sample of each speaker. We first transcribed every speech sample, in order to compare them to the listeners' transcription. Next, we calculated the percentage of correct words from each transcription to generate intelligibility scores for each speaker, following Derwing and Munro (2005). Note that each speaker received 14 intelligibility scores (one from each listener).

As for the comprehensibility ratings and the phonological control ratings, we simply built spreadsheets using the data file provided by Google Forms, organizing the rating for each speaker. Again, each speaker received 14 comprehensibility ratings (9-point scale) and 14 phonological control ratings (6-point scale). Although the phonological control scale con-

[3] The original scale was 1 'no difficulties' and 9 'extremely difficult'. We inverted the scale to make the correlational analysis easier to interpret. The formula used to invert the scale was adding the minimum (1) and maximum values (9) and then subtracting this value from each variable.

tained descriptors and levels (A1 to C2), for statistical purposes, we converted the descriptors into numbers ranging from 1 (A1) to 6 (C2).

The variables used to run the correlational analysis are (i) intelligibility scores, (ii) comprehensibility rates, and (iii) phonological control rates. Pearson Correlation was used to explore the relationship between variables. Significant correlations should be obtained if we find probability (p) values equal to or smaller than .016 (we applied Bonferroni correction by dividing the alpha level of .05 by 3, which is the number of correlation tests run in the study). Since this study contains data provided by raters, it is important to know if the raters agree and to what extent they agree. For this purpose, an inter-rater reliability test was conducted before calculating average scores and rates and running descriptive statistics for each variable. Reliability tests generate a Cronbach's alpha coefficient, which is an appropriate indicator of consistency across raters. Statistical analysis was carried out using JASP (0.19.3) and Excel.

4 Results and Discussion

The research question guiding the study is: How do raters' judgements of L2 learners' intelligibility and comprehensibility relate to the CEFR scale for phonological control? To analyze and interpret the results of the collected data, we first ran an inter-rater reliability test, Cronbach Alpha, to examine if the raters agreed with each other in their assessments. The Shapiro-Wilk normality test was also run to check for normal distribution in the data, as well as descriptive statistics. Finally, we ran Pearson correlation tests for the three variables (intelligibility, comprehensibility and phonological scale levels) to investigate the relationship between them. The results of these tests are going to be presented in this section, starting with Cronbach's interrater reliability analysis.

This section is going to be organized into three subsections. First, we report on the inter-rater reliability results and then move on to present the correlational analysis between the two speech dimensions and CEFR phonological control scale. The section ends with a discussion of the results of the correlational analysis.

4.1 Interrater reliability

Considering that the data of this study were provided by raters, it is necessary to examine if they agree with each other and to what extent (Larson-Hall, 2010). We conducted an inter-rater reliability test for each of the variables analyzed in this study: intelligibility, comprehensibility and phonological control. This type of statistical test allows comparing multiple ratings assigned to each participant and deciding whether all ratings can be averaged and used as a single score before running other statistical tests. A common test used for inter -rater reliability is called Cronbach Alpha, which is a measurement of intraclass correlation. This coefficient estimates the internal consistency between participants and items, and it is generally proposed that a value between 0.70–0.80 indicates a very good level of agreement between raters (Larson-Hall, 2010).

The Cronbach Alpha Analysis showed a high reliability rate for intelligibility (Cronbach $\alpha = .81$), comprehensibility ($\alpha = .85$) and for the phonological control variable ($\alpha = .87$). Thus, we concluded that the raters demonstrated high reliability levels, which allows us to calculate a mean rate for each of the variables in order to proceed with the statistical analysis. For reference, the original scores provided by the 14 raters are available in Martins (2022).

4.2 Relationship between the L2 speech dimensions and the CEFR phonological scale

Table 2 displays the descriptive statistics for this study. For the intelligibility score, the percentage for each participant was calculated by counting the number of correct words transcribed by the raters. As the results displayed in Table 2 show, the average minimum score for the intelligibility variable was 94.71 and the average maximum was 99.86, meaning that all participants were highly intelligible (m = 97.96, sd = 1.29).

For the second variable, comprehensibility, the raters used a scale ranging from 1 (extremely difficult to understand) to 9 (no difficulties to understand the speech). The average minimum rate assigned by the raters was 5.92 and the average maximum rate was 9 (m = 7.72, sd = .77); this means that the raters assigned mildly harsh comprehensibility rates even though they could understand the speech samples fairly well, as demonstrated by the high intelligibility scores.

Table 2 – Descriptive statistics for intelligibility, comprehensibility, and phonological control

N = 16	Mean	SD	Average Min.	Average Max.
Intelligibility (0-100)	97.96	1.29	94.71	99.85
Comprehensibility (1-9)	7.72	.77	5.92	9
Phonological Control (1-6)	3.46	.68	2.43	4.93

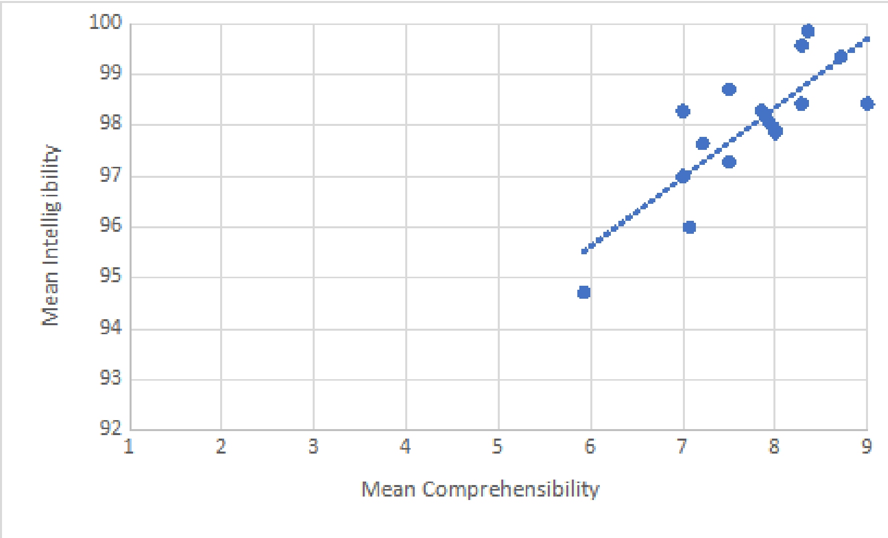
Min. and max. raw scores assigned: intelligibility = 86-100; comprehensibility = 2-9; Phonological control = 1-6

Source: The authors

Finally, the speakers' phonological control was assessed with the CEFR phonological scale. As previously explained, this scale has 6 descriptors, where 1 refers to A1 and 6 to C2. According to Table 2, the average minimum score assigned for phonological control was 2.43 and the average maximum was 4.93. The mean rate for phonological control (m = 3.46, sd = .68) is close to values within the B1 level descriptors ("Pronunciation is generally intelligible; intonation and stress at both utterance and word levels do not prevent understanding of the message. Accent is usually influenced by the other language(s) they speak.").

In order to answer the research question (How do raters' judgements of L2 learners' intelligibility and comprehensibility relate to the CEFR scale for phonological control?), we ran Pearson correlations using the three variables, namely, intelligibility, comprehensibility, and phonological control. We corrected the alpha value using the Bonferroni method, and a significant correlation should yield a p value equal to or smaller than .016. As Image 2 shows, there is a strong, positive and significant relationship between the intelligibility scores and the comprehensibility rates ($r(14) = .80$, $p = <.001$). This means that the more intelligible the speaker is, the easier it is to comprehend him/her.

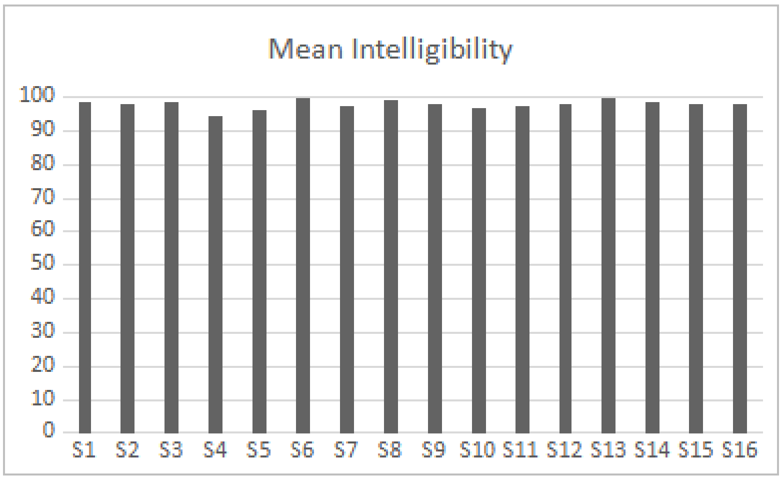
Image 2 – Correlation between intelligibility and comprehensibility



Source: The authors

Individual intelligibility scores for each speaker are shown in Image 3. This graph confirms that all speakers were highly intelligible, given that the listeners managed to transcribe over 90% of what they said. The lowest intelligibility score was 94%, obtained by S4.

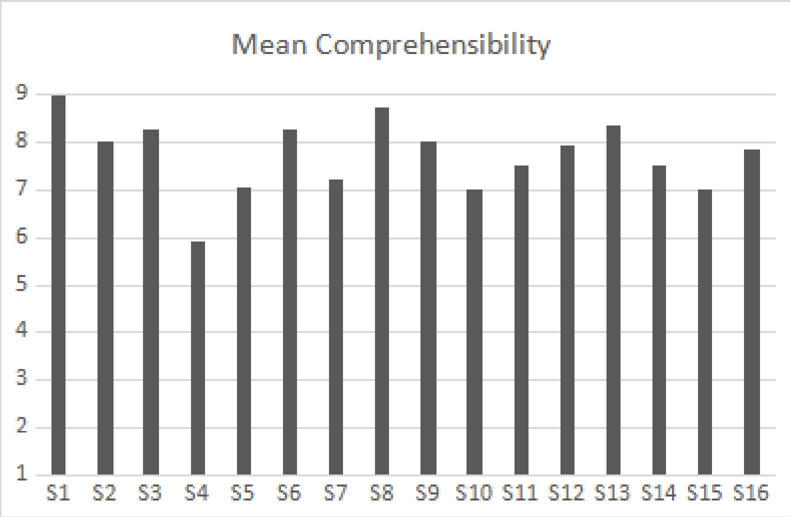
Image 3 – Listeners' mean intelligibility scores



Source: The authors (2025)

Turning to comprehensibility, individual speaker's rates are displayed in Image 4. One speaker (S1) was rated as highly comprehensible (comprehensibility rate = 9), meaning listeners had no difficulty at all understanding his/her speech. On the other hand, S4 was perceived as somewhat hard to comprehend, and this speaker is also the one who obtained the lowest intelligibility scores (5.92). Overall, the results in Image 4 show that the listeners assigned mildly harsh comprehensibility rates, despite being able to understand most of what the speakers said.

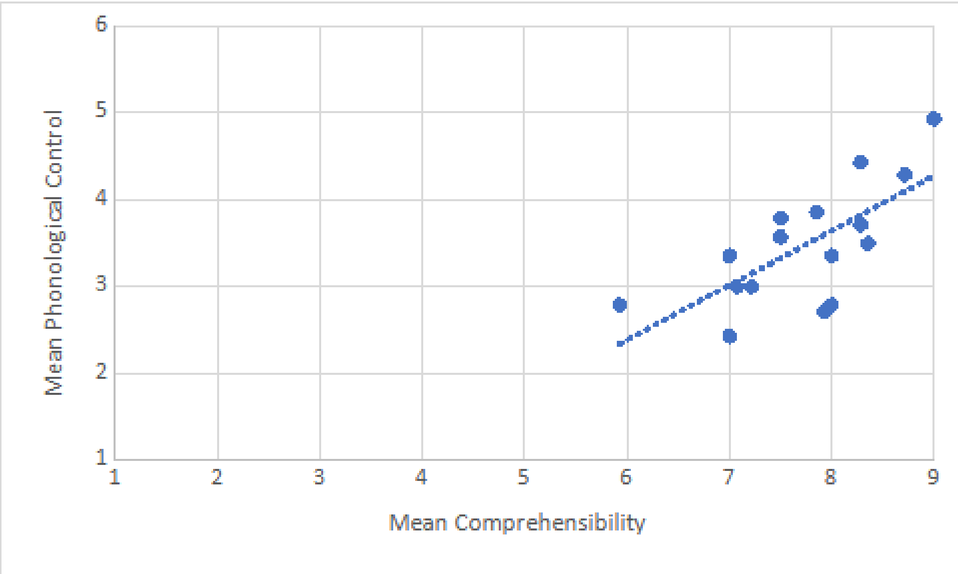
Image 4 – Listeners' mean comprehensibility rates



Source: The authors.

Examining the correlation results involving the phonological control variable, Image 5 also shows that there is a strong, positive, non-significant relationship between the average comprehensibility rates and the average phonological control rates ($r(14) = .69$, $p = .03$). This means that there is a tendency for speakers who are easier to comprehend to receive higher phonological control rates, yet this correlation did not reach statistical significance, considering the corrected p value adopted in this study.

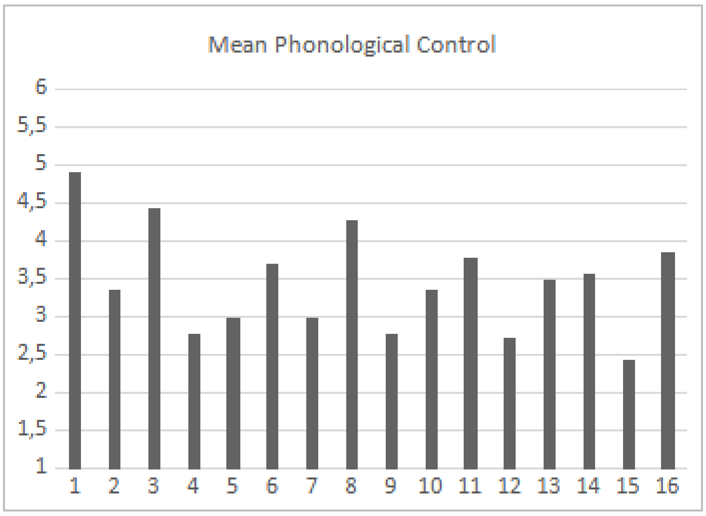
Image 5 – Correlation between comprehensibility and phonological control



Source: The authors.

Finally, as demonstrated in Image 6, there is a moderate, positive and non-significant relationship between the mean intelligibility score and the mean phonological control score ($r(14) = .42$, $p = .09$). This means that that there was a slight tendency for the more intelligible speakers to receive higher ratings from the listeners as regards phonological control. However, no statistical significance was obtained, meaning that neither intelligibility scores nor comprehensibility rates are reliable indicators of phonological control than intelligibility for this small group of speakers.

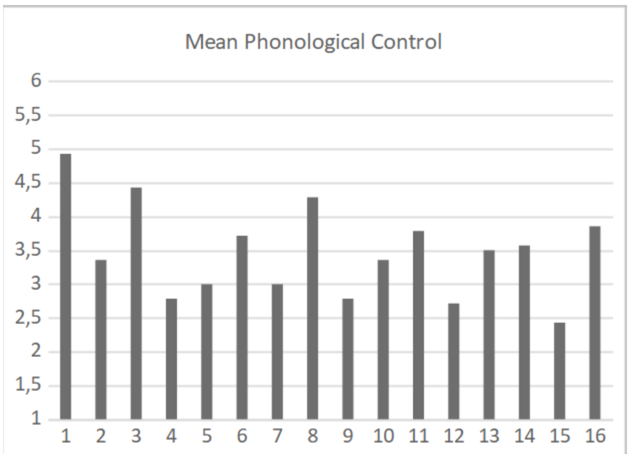
Image 6 – Correlation between intelligibility and phonological control



Source: The authors.

Turning to speakers' individual rates for phonological control, Image 7 shows a range of levels, with six speakers receiving a rate equal to or below 3, three speakers receiving rates above 4, and the remaining speakers receiving rates between 3-4. This indicates that the listeners considered that the speakers' level of phonological control was predominantly within the range of independent users (B1-B2).

Image 7 – Listeners' mean phonological control rates



Source: The authors.

The research hypothesis stated that there would be a positive, significant correlation between intelligibility, comprehensibility and phonological control, suggesting that the descriptors for phonological control yield results similar to those provided by measures of speech intelligibility and comprehensibility commonly used in L2 speech research, thus allowing adequate assessment of L2 pronunciation levels (A1-C2). Our hypothesis is partially confirmed by the correlational analysis results. Regarding the comprehensibility and the intelligibility variables, a significant correlation was found, but no significant correlation was found between the intelligibility and comprehensibility variables and the phonological control variable.

4.3 Summary and discussion

This study showed that there was a highly significant correlation between the intelligibility and comprehensibility variables, but neither intelligibility nor comprehensibility are significantly correlated with the phonological control variable. A moderate, non-significant correlation was obtained for intelligibility and phonological control, which suggests that comprehensibility is a better predictor of phonological control rates than intelligibility when experienced listeners (language teachers) assess L2 speech.

The hypothesis, which stated that there is a significant correlation between the intelligibility scores, comprehensibility rates and the phonological control rates was partially confirmed, considering that a significant relationship was not found between the phonological control levels assigned to speakers and the other speech dimension variables (i.e. intelligibility and comprehensibility).

Relating our findings with previous studies in the field of pronunciation assessment, similar to Saito *et al*. (2015), we found that the raters distinguished different levels when assigning comprehensibility rates, which is an indication that comprehensibility assessment is harsher than intelligibility assessment measured in the form of a transcription task (Derwing; Munro, 1997). In other words, even when listeners can understand most of what the speakers say (in the case of this study, over 90% of the speech samples were correctly transcribed), they are likely to identify pronunciation features (and possibly lexical and grammatical features) that impose some difficulty for listening comprehension.

It is interesting to notice that despite the low variability level in the intelligibility scores assigned by the listeners (all speakers received scores from 94% to 99%, thus showing a ceiling effect) a strong correlation between intelligibility and comprehensibility was found. This result corroborates previous studies that predict a strong relationship between these two variables, even though they seem to be associated with different speech features (Derwing; Munro, 1995b).

As for the CEFR phonological control scale, listeners also used it in a way that allowed them to distinguish different levels when rating the speakers. This result is aligned with what was observed for the comprehensibility variable. In other words, both the phonological control scale and the comprehensibility rating scale made it possible to assign different levels of performance for each speaker. Although the two variables yielded a strong correlation, no significant statistical level was obtained, possibly due to the small sample size, combined with the narrow range of comprehensibility rates (average rates ranged from 5 to 9). The correlation between phonological control and intelligibility variables was moderate and reached no significance either, thus implying that intelligibility was not a reliable indicator of phonological control for the participants of this study, given that even the speakers with lower levels of phonological control had highly intelligible speech. Once more the small sample size and the narrow range of intelligibility scores (average scores varied mostly between 94% to 99%) might have influenced the results.

Because the intelligibility variable did not present a range of scores, it is less useful to discriminate among different proficiency levels. In that sense, for assessment purposes, it seems that both the comprehensibility scale and the phonological control scale are more useful, as they allow distinguishing speakers across different levels of pronunciation performance. A possible reason for this finding may be that both scales allow raters to focus on specific pronunciation features. As Saito *et al.* (2015) explain, when assigning comprehension ratings to intermediate-to-advanced level learners, raters focus on segmental and suprasegmental features, which are also highlighted by the CEFR Phonological Control Scale descriptors. Finally, we can also speculate that the high intelligibility scores observed in the present study are partly due to the listeners' familiarity with the pronunciation features of English spoken by Brazilian users of English (Bent; Bradlow (2003); Silveira; Silva (2018); Kang et al. (2019)).

5 Conclusion

We could not obtain a significant correlation between phonological control and the comprehensibility and intelligibility variables, probably due to limited range of scores for the intelligibility variable and the narrow range of rates for comprehensibility. Apparently all 16 speakers displayed reasonably good command of spoken English to perform the image description task, which did not allow us to obtain a varied range of performances for the two speech dimension variables in this study. This problem of having a narrow range of scores was more clearly evidenced in the correlational analysis involving the intelligibility and the phonological control variables, which yielded a moderate, non-significant correlation. Conversely, comprehensibility and phonological control measures yielded a strong, non-significant correlation, possibly because the comprehensibility rates showed a slightly broader range than the intelligibility scores. Based on the correlational analysis results, we can speculate that the phonological control variable seems to assess a construct that is somehow different from the speech features assessed by the intelligibility and comprehensibility measures used in this study.

A remark often made by the raters during data collection was concerning the difficulty in using the phonological control scale without receiving proper training. Even though the raters were all experienced teachers of English as an L2 with background in assessment, the fact of not receiving specific training on the CEFR phonological control scale descriptors was an issue. Due to time restrictions, we could not offer training, and this was one of the limitations of this study; different results might be achieved by providing the participants with proper training sessions for them to get familiarized with the scale and to fully understand each descriptor, which is a suggestion for further research in the topic. Although we recognize the importance of training for increasing rater reliability (Kang *et al.*, 2019), we still managed to obtain high inter-rater reliability coefficients in our study, given that we selected experienced teachers to act as raters.

Replicating the study with a bigger population may as well result in different correlation coefficients between the three variables. Sound quality was also a factor that affected the ratings, and an important recommendation is to collect speech samples with the proper equipment in a lab; this was not possible for our study because of the Covid-19 pandemic and its restrictions. Nonetheless, these results must be taken with some caution, given that in the present study, both speakers and listeners share the same L1, which means that communication (intelligibility) may have been facilitated because the listeners (experienced English teachers) are familiar with the speakers' L2 pronunciation features. Finally, we need to acknowledge the limits of a correlational analysis. As pointed out by one of the anonymous reviewers, this type of analysis requires running multiple tests and, as we had multiple rates and scores for each speaker, we had to average the intelligibility scores and the comprehensibility and phonological control rates in order to run the correlational analysis. These procedures might have impacted the p values obtained for the three correlations somehow.

Intelligibility and comprehensibility were the main L2 speech dimensions examined in this research. The objective of the study was to understand the relationship between these two dimensions and the assessment of phonological control by raters using the CEFR (2018) scale. In this study, accentedness was not considered a relevant feature of speech to be taken into consideration for assessment purposes. The idea of working with the intelligibility and comprehensibility constructs, as mentioned in the introduction of this work, and later discussed in the literature review, was that assessment should focus on whether speakers of English as an L2 can sustain a conversation with their interlocutors and establish communication without demanding too much effort from listeners. With that in mind, this study is aligned with the idea of taking the constructs of intelligibility and comprehensibility to the classroom when teaching pronunciation, leaving accentedness aside. Focusing on intelligibility and comprehensibility can help with learners' confidence and performance in the language, as they do not need to be concerned with their accents or with sounding native-like.

The use of a specific scale to assess phonological control seems to be more aligned with what listeners intuitively do when they use a holistic scale to rate speech comprehensibility. This means that the CEFR phonological control scale can be an important instrument to assess L2 pronunciation and provide both teachers and learners with detailed information about which pronunciation features demand a lot of effort from their listeners to understand what they are trying to communicate. Based on this feedback, learners can seek strategies to modify their pronunciation in a way that allows them to communicate successfully and keep their listeners interested in interacting with them.

Statement of Authorship

This study reports on data from Thaisy da Silva Martins' MA thesis, which was supervised by Professor Rosane Silveira. Both authors worked together to design the experiment and complete the data analysis, including the statistical analysis. The first author was in charge of gathering data and transferring the data to spreadsheets for data analysis. The second author defined the scope of the article and wrote the first draft. Both authors collaborated on interpreting results, writing, and revising the article.

Acknowledgements

The authors would like to thank CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for the financial support, and the participants for contributing with this research.

References

ALLAN, D., Oxford Placement Test (1 or 2). Oxford: Oxford University Press, 2004.

BENT, T.; BRADLOW, A. R. The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America*, Melville, v. 114, p. 1600–1610, 2003.

COUNCIL OF EUROPE**.** *Common European Framework of Reference for Languages - Companion Volume.* 2018. Retrieved from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

COUNCIL OF EUROPE. *Common European framework of reference for languages: Learning, teaching, assessment***.** Cambridge: Press Syndicate of the University of Cambridge. 2001.

DERWING, T.; MUNRO, M. Accent, Intelligibility, And Comprehensibility: Evidence from Four L1s. *Studies in Second Language Acquisition,* Cambridge, v. 19, n.1, p. 1–16. 1997.

DERWING, T.; MUNRO, M. Second Language Accent and Pronunciation Teaching: A Research-Based Approach. *TESOL Quarterly,* Alexandria,v. 39, n.3, p. 379–397. 2005.

DERWING, T.; MUNRO, M. *Pronunciation fundamentals:* evidence-based perspectives for L2 teaching and research. Amsterdam: Benjamins. 2015.

DERWING, T., MUNRO, M.; THOMSON, R. A Longitudinal Study of ESL Learners' Fluency and Comprehensibility Development. *Applied Linguistics,* Oxford,v. 29, n.3, p. 359-380. 2008.

DEYGERS, B., ZEIDLER, B., VILCU, D.; CARLSEN, C. H. One Framework to Unite Them All? Use of the CEFR in European University Entrance Policies. *Language Assessment Quarterly,* Philadelphia,v. 15, n.1, p. 3–15. 2018.

FIGUERAS, N. The impact of the CEFR. *Elt Journal*, [S.L.], v. 66, n. 4, p. 477-485, 17 set. 2012. Oxford, Oxford University Press.

FOOTE, J. A.; TROFIMOVICH, P. Is it because of my language background? A study of language background influence on comprehensibility judgments. *Canadian Modern Language Review***,** Toronto, v. 74, n.2, p. 253–278. 2018.

HARDING, L. What do raters need in a pronunciation scale? The user's view. In ISAACS, T.; TROFIMOVICH, P. (eds) *Second Language Pronunciation Assessment*. Bristol: Multilingual Matters, 2017. p. 12-34.

ISAACS, T. Assessing pronunciation. *In* KUNNAN, A.J. (org.) *The Companion to Language Assessment.* Hoboken: Wiley, 2013. p. 140–155.

ISAACS, T.; THOMSON, R. Rater experience, rating scale length, and judgments of L2 pronunciation: revisiting research conventions. *Language Assessment Quarterly*, [*s. l*], v. 2, n. 10, p. 135-159, 2012.

KANG, O.; KERMAD, A. Assessment in second language pronunciation. *In* KANG, O. THOMSON, R.; MURPHY, J. (orgs.) *The Routledge Handbook of Contemporary English Pronunciation*. New York: Routledge, 2018. p. 511-526.

KANG, O., RUBIN, D; KERMAD, A. Effect of training and rater individual differences on oral proficiency assessment. *Language Testing*, London, v. 36, n.4, p. 481-504, 2019.

KANG, O., THOMSON, R. I.; MORAN, M. Empirical Approaches to Measuring the Intelligibility of Different Varieties of English in Predicting Listener Comprehension. *Language Learning*, Ann Arbor,v. 68, n.1, p. 115–146. 2017.

KHABBAZBASHI, N.; GALACZI, E. D. A comparison of holistic, analytic, and part marking models in speaking assessment. *Language Testing*, London, v. 37, n.3, p. 333–360. 2020.

KUIKEN, F., VEDDER, I. Rating written performance: What do raters do and why? *Language Testing*, London, v. 31, n.3, p. 329–348. 2014.

LARSON-HALL, J. *A Guide to Doing Statistics in Second Language Research Using SPSS*. New York: Routledge, 2010.

LEVIS, J. M. Changing Contexts and Shifting Paradigms in Pronunciation Teaching. *TESOL Quarterly*, Alexandria, v. 39, n.3, p. 369–377. 2005.

LEVIS, J. M. Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation*, Amsterdam,v. 6, n.3, p. 310-328. 2020.

MARTINS, T.S. *Using CEFR's phonological control scale to assess L2 learners' intelligibility and comprehensibility*. 2022. 95 f. Dissertação (Mestrado Letras Inglês) - Universidade Federal de Santa Catarina, Centro de Comunicação e Expressão, Programa de Pós-Graduação em Inglês: Estudos Linguísticos e Literários, Florianópolis, 2022. Disponível em: https://bu.ufsc.br/teses/PPGI0225-D.pdf. Acesso em: 12 maio 2023.

MUNRO, M. J.; DERWING, T. M. Processing Time, Accent, and Comprehensibility in the Perception of Native and Foreign-Accented Speech. *Language and Speech*, Thousand Oaks, v. 38, n.3, p. 289–306. 1995a.

MUNRO, M. J.; DERWING, T. M. Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, Ann Arbor,v. 45, n.1, p. 73–97. 1995b.

MUNRO, M. J.; DERWING, T. M. A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation*, Amsterdam,v. 1, n.1, p. 11–42. 2015.

QUEVEDO-CAMARGO, G. Breve história da evolução do construto proficiência em línguas. *Em Aberto*, Brasília, v. 32, n. 104, p. 27-44, 2019.

SAITO, K., TROFIMOVICH, P.; ISAACS, T. Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, Cambridge, v. 7, n.2, p. 217–240. 2015.

SILVEIRA, R.; MARTINS, T. S. Assessing second language oral proficiency development with holistic and analytic scales. *Ilha do Desterro*, Florianópolis, v. 73, p. 227-250, 2020.

SILVEIRA, R.; SILVA, T. C. L2 Speech intelligibility: effects of coda modification, degree of semantic information and listeners' background. *Revista Brasileira de Linguística Aplicada*, Belo Horizonte, v. 18, p. 639-664, 2018.

TOFFOLI, S.; ANDRADE, D.; BORNIA, A.C.; QUEVEDO-CAMARGO, G. Avaliação com itens abertos: validade, confiabilidade, comparabilidade e justiça. *Educação e Pesquisa*, São Paulo, v. 42, n. 2, p. 343-358, abr. 2016.

TOPAL, İ. H. CEFR-oriented probe into pronunciation: Implications for language learners and teachers. *Journal of Language and Linguistic Studies*, Konya,v. 15, n.2, p. 420–436. 2019

TROFIMOVICH, P.; ISAACS, T. Disentangling accent from comprehensibility. *Bilingualism,* Cambridge, v. 15, n.4, p. 905–916. 2012.

Annex 1 – CEFR Phonological control scale (2018 version)

	OVERALL PHONOLOGICAL CONTROL	SOUND ARTICULATION	PROSODIC FEATURES
C2	Can employ the full range of phonological features in the target language with a high level of control - including prosodic features such as word and sentence stress, rhythm and intonation - so that the finer points of their message are clear and precise. Intelligibility and effective conveyance and enhancement of meaning are not affected in any way by features of accent that may be retained from other language(s).	Can articulate virtually all the sounds of the target language with clarity and precision.	Can exploit prosodic features (e.g. stress, rhythm and intonation) appropriately and effectively in order to convey finer shades of meaning (e.g. to differentiate and emphasise).
C1	Can employ the full range of phonological features in the target language with sufficient control to ensure intelligibility throughout. Can articulate virtually all the sounds of the target language; some features of accent(s) retained from other language(s) may be noticeable, but they do not affect intelligibility.	Can articulate virtually all the sounds of the target language with a high degree of control. They can usually self-correct if they noticeably mispronounce a sound.	Can produce smooth, intelligible spoken discourse with only occasional lapses in control of stress, rhythm and/or intonation, which do not affect intelligibility or effectiveness. Can vary intonation and place stress correctly in order to express precisely what they mean to say.
B2	Can generally use appropriate intonation, place stress correctly and articulate individual sounds clearly; accent tends to be influenced by the other language(s) they speake, but has little or no effect on intelligibility.	Can articulate a high proportion of the sounds in the target language clearly in extended stretches of production; is intelligible throughout, despite a few systematic mispronunciations. Can generalise from their repertoire to predict the phonological features of most unfamiliar words (e.g. word stress) with reasonable accuracy (e.g. while reading).	Can employ prosodic features (e.g. stress, intonation, rhythm) to support the message they intend to convey, though with some influence from the other languages they speak.
B1	Pronunciation is generally intelligible; intonation and stress at both utterance and word levels do not prevent understanding of the message. Accent is usually influenced by the other language(s) they speak.	Is generally intelligible throughout, despite regular mispronunciation of individual sounds and words they are less familiar with.	Can convey their message in an intelligible way in spite of a strong influence on stress, intonation and/or rhythm from the other language(s) they speak.
A2	Pronunciation is generally clear enough to be understood, but conversational partners will need to ask for repetition from time to time. A strong influence from the other language(s) they speak on stress, rhythm and intonation may affect intelligibility, requiring collaboration from interlocutors. Nevertheless, pronunciation of familiar words is clear.	Pronunciation is generally intelligible when communicating in simple everyday situations, provided the interlocutor makes an effort to understand specific sounds. Systematic mispronunciation of phonemes does not hinder intelligibility, provided the interlocutor makes an effort to recognise and adjust to the influence of the speaker's language background on pronunciation.	Can use the prosodic features of everyday words and phrases intelligibly, in spite of a strong influence on stress, intonation and/or rhythm from the other languages(s) they speak. Prosodic features (e.g. word stress) are adequate for familiar everyday words and simple utterances.
A1	Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by interlocutors used to dealing with speakers of the language group. Can reproduce correctly a limited range of sounds as well as stress for simple, familiar words and phrases.	Can reproduce sounds in the target language if carefully guided. Can articulate a limited number of sounds, so that speech is only intelligible if the interlocutor provides support (e.g. by repeating correctly and by eliciting repetition of new sounds).	Can use the prosodic features of a limited repertoire of simple words and phrases intelligibly, in spite of a very strong influence on stress, rhythm and/or intonation from the other language(s) they speak; their interlocutor needs to be collaborative.

Source: Adapted from CEFR Companion Volume (Council of Europe, 2018)