



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE COMUNICAÇÃO E EXPRESSÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INGLÊS: ESTUDOS LINGUÍSTICOS E
LITERÁRIOS

Thaisy da Silva Martins

**Using CEFR's Phonological Control Scale to Assess L2 Learners' Intelligibility and
Comprehensibility**

FLORIANÓPOLIS

2022

Thaisy da Silva Martins

**Using CEFR's Phonological Control Scale to Assess L2 Learners' Intelligibility and
Comprehensibility**

Dissertação submetida ao Programa de
Pós-graduação em Inglês: Estudos Linguísticos e
Literários da Universidade Federal de Santa
Catarina para a obtenção do título de Mestre em
Inglês: Estudos Linguísticos e Literários
Orientador: Prof. Dr. Rosane Silveira

Florianópolis

2022

Ficha de identificação da obra

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Martins, Thaisy
Using CEFR's phonological control scale to assess L2
learners' intelligibility and comprehensibility / Thaisy
Martins ; orientadora, Rosane Silveira, 2023.
96 p.

Dissertação (mestrado) - Universidade Federal de Santa
Catarina, Centro de Comunicação e Expressão, Programa de Pós-
Graduação em Inglês: Estudos Linguísticos e Literários,
Florianópolis, 2023.

Inclui referências.

1. Inglês: Estudos Linguísticos e Literários. 2.
Inteligibilidade. 3. Compreensibilidade. 4. Avaliação. 5.
CEFR. I. Silveira, Rosane. II. Universidade Federal de
Santa Catarina. Programa de Pós-Graduação em Inglês: Estudos
Linguísticos e Literários. III. Título.

Thaisy da Silva Martins

Using CEFR's Phonological Control Scale to Assess L2 Learners' Intelligibility and Comprehensibility

O presente trabalho em nível de Mestrado foi avaliado e aprovado, em 16 de dezembro de 2022, pela banca examinadora composta pelos seguintes membros:

Prof.(a) Gladys Pens de Quevedo Pereira de Camargo, Dr.(a)

Universidade de Brasília (UnB)

Prof.(a) Hanna Kivisto de Souza, Dr.(a)

Universidade Federal de Santa Catarina (UFSC)

Prof.(a) Donesca Puntel, Dr.(a)

Universidade Federal de Santa Catarina (UFSC)

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestre em Inglês: Estudos Linguísticos e Literários.

Coordenação do Programa de Pós-Graduação

Prof.(a) Rosane Silveira, Dr.(a)

Orientador(a)

Florianópolis, 2023.

ACKNOWLEDGEMENTS

My first thank you goes to my advisor, Professor Rosane Silveira, for guiding me through this journey. Thank you for being such a kind person, for your support and for teaching me so much all these years, since I was an undergraduate student taking my first steps in scientific research. You are the greatest advisor I could ever have.

I would like to express my infinite gratitude to my parents, Fátima and José, for encouraging me to study since I was a very young girl, and to pursue my dreams. I know how difficult it was for you when I left home to do that. Thank you to my big sister, Janini, for being my best friend in life and for always being there for me. You are my favorite person.

A warm thank you to my partner, João Emílio, for being by my side, for never once doubting that I could achieve anything, and for thinking I am the most amazing person alive. I admire you that much as well.

Thank you to all of my friends who helped me in this process, each in their own way. To Sara and Lia, for showing me how sometimes my anxiety was absurd and making me laugh about it; you girls made everything feel lighter. A special thanks to João Luiz and Fernanda; this experience would not have been the same without you two, and I'm glad we decided to walk this path together.

I am very grateful for the financial support provided by CAPES, and for everybody at PPGI. Finally, I would like to thank the members of the examining committee, Professor Gladys Plens de Quevedo Pereira de Camargo, Professor Hanna Kivisto de Souza, and Professor Donesca Puntel, for kindly accepting to read and contribute to my work.

ABSTRACT

The objective of this study is to examine the adequacy of the new CEFR phonological scale as a tool to assess speech intelligibility and comprehensibility. In order to do that, we collected data from 16 speakers, who were recorded describing an image in English. Their speech samples were assessed by 14 listeners, teachers of English as a second language, in terms of intelligibility, comprehensibility and phonological control. The raters transcribed the speech samples of the participants, in order for us to generate an intelligibility score; for comprehensibility, the raters assigned a level of difficulty in understanding the speech through a Likert scale; finally, they used the CEFR scale for phonological control to designate a level (A1-C2) for each speaker. We ran Pearson correlation to investigate the relationship between the variables intelligibility and comprehensibility; intelligibility and phonological control; and comprehensibility and phonological control. Results demonstrate that there is a highly significant correlation between intelligibility and comprehensibility, and between comprehensibility and phonological control as well. The raters also responded to a questionnaire with open and closed questions to report their experience using the CEFR phonological control scale. The results of the survey suggest that the scale is a useful tool to assess intelligibility and comprehensibility, but there is room for improvement, as well as a proper training for raters to use the scale.

Key-words: Intelligibility. Comprehensibility. Assessment.

RESUMO

O objetivo deste estudo é examinar a adequação da nova escala fonológica do CEFR como instrumento de avaliação da inteligibilidade e compreensibilidade da fala. Para isso, coletamos dados de 16 falantes, que foram gravados descrevendo uma imagem em inglês. Suas amostras de fala foram avaliadas por 14 ouvintes, professores de inglês como segunda língua, quanto à inteligibilidade, compreensibilidade e controle fonológico. Os avaliadores transcreveram as amostras de fala dos participantes, para gerarmos um escore de inteligibilidade; para compreensibilidade, os avaliadores atribuíram um nível de dificuldade na compreensão da fala por meio de uma escala Likert; por fim, utilizaram a escala CEFR para controle fonológico para designar um nível (A1-C2) para cada falante. Calculamos o coeficiente de correlação de Pearson para investigar a relação entre as variáveis inteligibilidade e compreensibilidade; inteligibilidade e controle fonológico; e compreensibilidade e controle fonológico. Os resultados demonstram que existe uma correlação altamente significativa entre inteligibilidade e compreensibilidade, e também entre compreensibilidade e controle fonológico. Os avaliadores também responderam a um questionário com perguntas abertas e fechadas para relatar sua experiência com o uso da escala de controle fonológico CEFR. Os resultados da pesquisa sugerem que a escala é uma ferramenta útil para avaliar a inteligibilidade e a compreensibilidade, com ressalvas quanto à necessidade de ajustes e treinamento dos avaliadores para o uso da escala.

Palavras-chave: Inteligibilidade. Compreensibilidade. Avaliação.

LIST OF FIGURES

Figure 1: Oxford Placement Test online version extract	32
Figure 2: Speech Elicitation Image	33
Figure 3: Listening Tasks	35
Figure 4: Sample of the questionnaire to report the experience using the phonological control scale	36
Figure 5: Standard deviation for the intelligibility variable	44
Figure 6: Standard deviation for the comprehensibility variable	44
Figure 7: Standard deviation for the phonological control variable	45
Figure 8: Correlation between intelligibility and comprehensibility	46
Figure 9: Correlation between comprehensibility and phonological control	47
Figure 10: Correlation between intelligibility and phonological control	48
Figure 11: Answers for question 1	54
Figure 11: Answers for question 2	55
Figure 11: Answers for question 3	56
Figure 11: Answers for question 4	56

LIST OF TABLES

Table 1 - CEFR Phonological control scale	22
Table 2 - Participants: speakers	27
Table 3 - Participants: listeners	29
Table 4 - Overall scores for intelligibility, comprehensibility, and CEFR Phonological Scale (combined means for the 14 listeners).	41
Table 5 - Pearson Correlation	45
Table 6 - Scores assigned for intelligibility (percentage of words transcribed correctly)	49
Table 7 - Comprehensibility scores assigned by the raters	50
Table 8 - Phonological Scale CEFR levels assigned by the raters	51

TABLE OF CONTENTS

1. INTRODUCTION	11
1.1 Objectives	11
1.2 Research Questions and Hypotheses	11
1.3 Significance of the Study	12
1.4 Organization of the Thesis	12
2. REVIEW OF LITERATURE	13
2.1 Intelligibility	13
2.2 Comprehensibility	15
2.3 Accent	17
2.4 Pronunciation Assessment	18
2.5 Raters and Rating Scales	19
2.6 CEFR	20
2.6.1 The CEFR levels and scales	20
2.6.2 Studies about the CEFR	24
3. METHOD	25
3.1 Participants	27
3.2 Instruments and Materials	29
3.2.1 Background Questionnaires	29
3.2.2 Consent Forms	30
3.3.3 Proficiency Test	30
3.3.4 Speech Elicitation	31
3.3.5 Listening Tasks	32
3.4 Procedures For Data Collection	34
3.4.1 Intelligibility Assessment	37
3.4.2 Comprehensibility Assessment	37
3.4.3 Assigning oral proficiency levels with the phonological control scale	38
3.5 Procedures for Data Analysis	39
4. RESULTS AND DISCUSSION	40
4.1 Introduction	40
4.2 Correlational Analysis Results	41
4.2.1 Interrater Reliability	41
4.2.2 Descriptive Statistics	42
4.2.3 Correlational Analysis	44
4.2.4 Individual results	49
4.2.3 Summary	53

	10
4.3 Survey Results	54
4.4 Summary	62
5. CONCLUSION	63
5.1 Limitations of the study and suggestion for further research	64
5.2 Pedagogical implications	65
REFERENCES	66
APPENDICES	69

1. INTRODUCTION

Assessment is a relevant part in the process of learning a second language. 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, that is, the repetition and imitation of vowel and consonant sounds in order to sound native. With time and understanding that the native-like pronunciation was not an attainable goal, pronunciation assessment started to take different paths, especially with the nativeness versus intelligibility principle proposed by Levis (2005).

According to Levis (2005), 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". That is, having a strong accent does not mean that the learner will not be understood by their interlocutor; the intelligibility concept holds that, if communication is established, there is no need for the learner 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 Munro and Derwing (1995a, p. 291), comprehensibility refers to "listeners' perception of difficulty in understanding particular utterances", while accentedness is defined as "how strong the talker's foreign accent is perceived to be".

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. Among other materials, it features descriptor scales that can be helpful when assessing L2 learners' proficiency. Recently, the CEFR was updated to include a new scale for phonological control, used for the assessment of oral proficiency in L2 learners. 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).

1.1 Objectives

Considering the new CEFR scale for phonological control, the concepts of intelligibility, comprehensibility and the context of Brazilian Portuguese learners of English as an L2, the objective of this study is to examine the adequacy of the new CEFR phonological scale as a tool to assess speech intelligibility and comprehensibility. The specific objectives are (a) to investigate how the intelligibility and comprehensibility constructs are incorporated in the CEFR phonological control scale, (b) to compare the performance of experienced raters when using measures of intelligibility and comprehensibility and the CEFR scale, and (c) to correlate the CEFR phonological control scale with intelligibility and comprehensibility measures commonly used by researchers.

1.2 Research Questions and Hypotheses

The study will try to answer the following research questions:

1. How do raters' judgements of L2 learners' intelligibility and comprehensibility relate to the new CEFR scale for phonological control?
2. To what extent is the improved CEFR scale helpful in the assessment of intelligibility by experienced L2 teacher raters?

Furthermore, the study has the following hypothesis:

1. There is a positive correlation between intelligibility measures, comprehensibility ratings and raters' responses of their experience using the CEFR scale, suggesting that the new descriptors for phonological control adequately reflect the results of intelligibility assessment for each proficiency level (A1-C2)

1.3 Significance of the Study

Having in mind that the new CEFR scale for phonological control was released in 2018, there has not been much research on the subject (TOPAL, 2019; KHABBAZBASHI; GALACZI, 2020) which is a motivation for me to contribute to the field. Seeing that this is a valuable resource in the assessment of L2 learners, as well as a high-stake document used in language assessment, I consider it important to examine the adequacy of the new descriptor

scale and relate them to the context of Brazilian Portuguese learners of English as an L2, students of the *Letras Inglês* program at UFSC.

As potential study-abroad and work-abroad candidates, taking standardized tests that evaluate proficiency tends to be common in the lives of these students. Since the CEFR scales are used as a reference to assess proficiency in these tests, I think it is relevant to examine this recently revised resource to learners that face different types of assessment throughout their education and afterwards as well. Besides, I intend to see how the new description scale for all levels of proficiency are helpful to the assessment of intelligibility in the educational context, in order to see if it can also help teachers in the development of non-standardized evaluations.

1.4 Organization of the Thesis

This research project is organized in the following way: first, the introduction contextualizing the study and stating the objectives, research questions, hypotheses and significance; after that, the method describing the participants and instruments for data collection as well as for data analysis; finally, a subsection demonstrating the probable contents of the MA thesis as well as a timetable of the research conduction.

2. REVIEW OF LITERATURE

In this section I will discuss the topics relevant to this study. I will present an overview of the studies in the area of intelligibility and comprehensibility, accent, pronunciation assessment, raters and rating scales. After that, I will present the Common European Framework levels and scales, and finally discuss studies related to the CEFR.

2.1 Intelligibility

A definition of intelligibility accepted and used in the literature is the one provided by Derwing and Munro (1995a, p. 291) “the extent to which an utterance is actually understood” and it can be measured through orthographic transcription of a speech sample, according to the authors. Derwing and Munro (1995b) analyzed accent¹, comprehensibility and intelligibility in the speech of second language learners. Native speakers of English judged these aspects on the speech of 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) that communication can be intelligible even when the speaker has a strong foreign accent.

Bent and Bradlow (2003) investigated the influence of the native language background on intelligibility. Native speakers of Chinese, Korean and English performed a task of sentence reading; then, native listeners of Chinese, Korean and English and a mixed group of listeners from different backgrounds listened to the sentences uttered by the speakers, and were asked to write down all the words they heard. The results demonstrated that native listeners considered native speakers more intelligible than non-native speakers; moreover, the “matched interlanguage speech intelligibility benefit” could be seen in the results. This benefit means that the non-native speech is more intelligible to the non-native listener due to the fact that they both share the same native language and this shared knowledge impacts intelligibility (BENT; BRADLOW, 2003).

An empirical study conducted by Kang, Thomson, and Moran (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 TOEFL. The participants were 18 high-proficient English users for the speaking test, and 60 listeners for the intelligibility and

¹ A listener’s perception of how different a speaker’s accent is from that of the L1 community (Derwing and Munro, 2005).

listening comprehension tests. The listeners assessed the speakers intelligibility through five different measures: true/false statements, scalar ratings, perception of nonsense sentences, perception of filtered sentences and transcription. The five measures were analyzed and the authors could conclude that the perception of nonsense sentences was effective in measuring the intelligibility of speakers from different backgrounds. A point that was highlighted was that the participants of the test had the same proficiency level (high) and that using participants of different levels could lead to more wide results; a positive characteristic of this study was that the authors used five different intelligibility measures, while most works in the area use only one or two (KANG; THOMSON; MORAN, 2017).

A study conducted by Derwing and Munro (2007) explored the relationship between intelligibility, comprehensibility and accentedness. The speech of speakers from four different L1s (Cantonese, Japanese, Polish, and Spanish) were evaluated by native speakers of English. The results suggest that the three dimensions investigated are related but not equivalent, meaning that speech samples evaluated as strongly accented were also perceived as highly intelligible. This research contributes significantly for the establishment of intelligibility as an achievable and sufficient goal for the L2 learner. It is an important addition for the development of the area and the understanding that, while intelligibility and accentedness are related constructs, they are also independent from each other.

A recent study conducted by Delatorre (2017) contributes to research on intelligibility in Brazil. In her study, the author investigated the intelligibility of regular verbs in the simple past tense. The listeners were 14 Brazilian learners of English who examined the intelligibility of the speakers through orthographic transcription. The results suggest that verb-familiarity, language experience and language proficiency correlate with verb intelligibility; besides, they also indicate that Brazilian listeners demonstrated some difficulty in the recognition of regular verbs in the simple past tense when inserted in short sentences. Another important finding is regarding the speakers and listeners L1: the listeners rated non-native speakers of English as more intelligible than native speakers of English, suggesting an interlanguage speech intelligibility benefit. This study is relevant for this research because it focuses on the Brazilian context, and, in comparison to most of the previously reviewed works, it has L2 English users as listeners. Both characteristics are going to be present in this research.

2.2 Comprehensibility

Comprehensibility is another important concept in the assessment of pronunciation, and it is often explored alongside intelligibility and accentedness. This concept, as defined by Derwing and Munro (1995a), refers to the listeners' perception of the difficulty in understanding speech samples; comprehensibility is usually measured through scalar ratings, such as Likert-scales. In the same study, the authors investigated the relationship between speech processing time (measured in terms of how long it took listeners to respond), judgments for accentedness and judgments for comprehensibility, provided by English native speakers. The results also support the idea that accent and comprehensibility are related, but not influenced by each other. Moreover, 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.

A study by Trofimovich and Isaacs (2012) investigated what linguistic features are related to accent and what features are related to comprehensibility, with the objective of determining which linguistic aspects are related to each dimension of speech. The speech production was provided by 40 French native speakers and evaluated by 60 inexperienced raters and three experienced teachers of English. The linguistic aspects investigated were phonology, fluency, lexis/grammar, and discourse. 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 aspects of speech are considered relevant when evaluating a speaker's accent. Regarding comprehensibility, the authors highlight the results that listeners tend to perceive grammatical errors as distractors for the comprehensible speech, contributing to a more effortful comprehension.

Regarding possible language background influences on comprehensibility judgments, Foote and Trofimovich (2018) investigated the role of the listeners native language in the process of comprehensibility judgements. The listeners were 40 L2 English speakers from Mandarin, French, Hindi, and English backgrounds. They rated speech samples of 30 L2 English speakers from Mandarin, French and Hindi backgrounds. The results demonstrated, in the verbal reports provided by the listeners and in the statistical analyses, 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) investigated the correlation between comprehensibility and accentedness for learners from different ability levels. In their study, 120 Japanese speakers of beginner, intermediate and advanced levels of English participated in a speech elicitation task of image description. Their speech was assessed by 5 inexperienced native speakers of English in the domains of comprehensibility and accentedness. Five native speakers of English and experienced raters evaluated the speech samples focusing on linguistic analyses of phonological, lexical and grammatical characteristics of speech. The results are in line with previous work in the area concerning 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 for each proficiency level, 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.

2.3 Accent

Accent is present and almost always inevitable, according to Derwing and Munro (1995), even for those who spend years in the L2 country. Accentedness is defined by the authors as a listener's perception of how different a speaker's accent is from that of the L1 community. Having a strong accent may result in miscommunication, due to the speaker's pronunciation of segmental aspects (vowels and consonants) or suprasegmental aspects (such as stress or intonation), and it might create confusion, irritation, or even prejudice against the speaker (DERWING; MUNRO, 1995). Derwing and Munro (1997) state that accent, intelligibility and comprehensibility are related but independent constructs, which demonstrate that these conclusions have been in the literature for quite some time. This relationship was investigated by the authors, and it was concluded that accentedness ratings are harsher than the comprehensibility ratings, which in turn are harsher than intelligibility scores. This suggests that communication was established, due to the positive ratings for

intelligibility and comprehensibility, even if some features of accent were perceived as strong (DERWING; MUNRO, 1997).

More recent studies in the area have explored which features are connect to these dimensions; Saito, Trofimovich and Isaacs (2015) mention that comprehensibility was linked to dimensions such as grammar, lexicon, pronunciation and discourse structure, while accentedness is mainly associated to segmental and suprasegmental features of speech. In this study, 120 Japanese speakers were rated by native speakers; their findings corroborate previous work in the area in the sense that, when rating comprehensibility, the listeners associated the speakers' performance with different linguistic aspects, while for accentedness, the attention was given to segmental accuracy. The authors highlight that an adult L2 speaker can achieve native-like proficiency in terms of vocabulary and grammar, while still being identified as non-native when speaking (SAITO; TROFIMOVICH; ISAACS, 2015).

In order to understand better the differences in the rating of comprehensibility and accentedness, Trofimovich and Isaacs (2012) investigated which linguistic features were taking into consideration when rating these two dimensions. The results were in-line with their expectations: grammar, vocabulary, and prosody are associated with comprehensibility, while segmental accuracy is often the main feature related to accented speech. Recent works in the area seem to have established that, while accent is a characteristic that will always be present in the speech of L2 speakers, even if they are highly proficient and lived for a long time at the L2 country, it is accepted, at least among linguists and students of the area, that it does not affect communication.

2.4 Pronunciation Assessment

Pronunciation assessment can vary in method and type of measurement. Kang and Kermad (2018) demonstrated that pronunciation can be measured either by human beings or by machines. The assessment provided by listeners/raters can use the method of rating scales, transcription, true/false questions or comprehension questions, for example. Although having listeners to evaluate is the most common way of measuring pronunciation, human raters can be biased due to a variety of reasons, which can impact in the process of evaluation and thus must be taken into account (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, Rubin and Kermad (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 82 naive raters, that is, they did not have any formal experience in evaluating oral proficiency, and they all varied in language background. The 82 raters holistically evaluated 112 speech samples produced by TOEFL examinees. The results demonstrated that English native-speakers tended to be less severe than non-native speakers when judging, even though the authors do not speculate a reason for that. Moreover, the authors argue that familiarity played a role in affecting the raters' judgement; the frequent contact with a certain accent impacted the raters in the sense of being more tolerant towards it (KANG; RUBIN; KERMADE, 2019). Another result demonstrated that a training session - which happened online - helped to balance the assessment, that is, after the training the raters' responses tended to be more homogenized (KANG; RUBIN; KERMADE 2019).

2.5 Raters and Rating Scales

Rating scales to assess oral proficiency, especially comprehensibility and accentedness, are used by human raters in order to provide a measurement of the understanding. Since this type of assessment is based on a listener's perception and judgment, many factors can influence the outcome of this rating, such as the speaker's proficiency level, type of task, topic of the assignment, rating experience, familiarity with rating scales, or training, (KUIKEN; VERDDEN, 2014) for example, as well as the raters' strictness, the difficulty of the task, the topic, the criteria and the rating scale (TOFFOLI; ANDRADE; BORNIA; QUEVEDO-CAMARGO, 2016). A rating scale is a sort of framework that enables a listener to judge an aspect of language in 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), nine-levels scales are often chosen due to its versatility, meaning that it can be used with learners from any L1, as well as it can be used by inexperienced raters who have no background in linguistics. The authors also highlight that Cronbach's alpha coefficients

(interrater reliability) are often high when using this method. This refers to interrater reliability, an important validity method when using human raters to assess speech, considering that there must be an agreement between the raters in order for their judgment to be reliable. In the same study, the authors investigated possible different outcomes when using a five-point Likert scale and a nine-point Likert scale. They did not find a ceiling effect, but rather concluded that a floor effect might have happened, due to the low proficiency of the participants; besides, they mention that some participants had some hesitancy in assigning level 1 for the speakers, due to the “moral failing attached to the low end of the scale” (ISAACS; THOMSON, 2012). Moreover, the authors present an interesting point for the discussion of five-point or nine-point rating scales: 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 (ISAACS; THOMSON, 2012). This suggests that we cannot suppose what raters would assign to different speech samples depending on the samples provided (for instance, a rater would assign a 6 in a 9-point scale for a speaker in one study and a 3 for the same speaker in a different study, because the other speakers were less accented in the second study) without standardizing the criteria between the two studies first.

The debate between expert raters vs. naive raters has different conclusions. Isaacs (2013), discusses that having naive raters, or laypeople, to assess learners’ speech may be interesting due to the fact that 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 due to the fact that naive raters have demonstrated to produce reliable judgments of L2 learners’ 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.

2.6 CEFR

This section will present the Common European Framework objectives, scales and descriptors, focusing on the phonological control scale, as well as some studies in the literature using the CEFR

2.6.1 The CEFR levels and scales

The Common European Framework of Reference for Languages is a material for language learning, teaching and assessment. It is a source for the development of language syllabuses, curriculum guidelines, examinations, textbooks and so on across Europe. The framework also provides descriptors for levels of proficiency and for assessment. The approach adopted by the CEFR comprises 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). This document had a huge impact on language learning and teaching, on the development of textbooks and high-stakes 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 operate and apply these constructs (QUEVEDO-CARMARGO, 2019).

The Common Reference Levels were developed to help on the description and measurement of levels of proficiency, creating a scheme to help in the comparison between different systems and standard language tests. The reference levels provided by the CEFR are breakthrough and waystage for the basic user (A), threshold and vantage for the independent user (B) and effective operational proficiency and mastery for the proficient user (C). For better contextualization of the levels, the description for the proficient user (C2) is:

“can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations”

while the basic user (A1) is described as

“can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.” (COUNCIL OF EUROPE, 2001).

The illustrative descriptor scales comprises 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 research is the phonological control scale. The description for phonological competence in the 2001 text 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).

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 level is described as “as C1” which, in its turn, is “can vary intonation and place sentence stress correctly in order to express finer shades of meaning”. These descriptions are quite vague and can lead to misplacement of candidates.

These descriptors were updated in 2018 in order to embrace more concepts related to oral proficiency and to provide more detailed guidelines for assessment (see Table 1). In the 2018 Companion Volume, there can be seen the integration of the articulation, prosody, accentedness and intelligibility constructs. Moreover, the 2018 version included descriptors for mediation, sign language, language learning for children, as well as presented three new sublevels: A1+, B1+ and B2+ (QUEVEDO-CAMARGO, 2019). The new descriptors for A1 and C2 level are, respectively:

“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 concerned. Can reproduce correctly a limited range of sounds as well as the stress on simple, familiar words and phrases.”

and C2 as

“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 his/her message are clear and

precise. Intelligibility and effective conveyance of and enhancement of meaning are not affected in any way by features of accent that may be retained from other language(s)". (COUNCIL OF EUROPE, 2018, p. 136).

An improvement can be seen in the update version that is more in-line with current research and studies in the area of oral proficiency. A new companion volume was published in 2020 with some alteration in the scales, however, the phonological control scale remained the same as the 2018 version.

Table 1

CEFR Phonological control scale

	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 speak, 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 employ prosodic features (e.g. stress, intonation, rhythm) to support the message they intend to convey, though with some influence from the

		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).	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.	<p>Pronunciation is generally intelligible when communicating in simple everyday situations, provided the interlocutor makes an effort to understand specific sounds.</p> <p>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.</p>	<p>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 language(s) they speak.</p> <p>Prosodic features (e.g. word stress) are adequate for familiar everyday words and simple utterances.</p>
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.	<p>Can reproduce sounds in the target language if carefully guided.</p> <p>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).</p>	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: CEFR Companion Volume (2018)

2.6.2 Studies about the CEFR

The Common European Framework of Reference for Languages provides linguistic competence scales to assess learners' proficiency, such as grammatical accuracy, vocabulary control, and phonological control. The focus of this study is on the phonological control scale, which has been recently updated to encompass more features. The 2001 scale

reinforced the idea of a desirable native-like accent and focused on the accuracy of segments. The 2018 version gives special attention to intelligibility, sound articulation and prosody, diminishing the focus on accuracy and accentedness (COUNCIL OF EUROPE, 2018).

Deygers et. al (2017) 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, 2017, p. 10). This article illustrates the importance and impact of the CEFR while also discussing its potential misuse.

A study conducted by Deygers and Van Gorp (2015) explored the possibility of creating a reliable CEFR-based assessment scale. A rating scale was constructed by novice raters prior to the study, and the researchers aimed at analyzing the usage of this scale having the same raters judging the speech of 200 participants. The results indicate that there was a high rate of inter-rater reliability, meaning an agreement among the listeners, which corroborates to a reliable use of the scales; however, the authors highlight that the scales could have been interpreted differently by the raters, suggesting a need for specific rater training before the use.

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 on language learning, teaching and assessment, especially on categorizing 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). Besides that, the language descriptors were also rapidly adopted, especially to aid in the development of language learning programmes, 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 determined 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, for example, for first language learning and teaching (FIGUERAS, 2012).

This review demonstrates that the Common European Framework can be misused and misinterpreted; the main objective of the framework is to serve as a reference for language teaching, learning and assessment, as its own name suggests (COUNCIL OF EUROPE, 2018), and it is important to adapt it to one's needs. It is important to consider the implications of using the assessment scales and what types of negative results they might bring, especially because the update of including intelligibility and comprehensibility constructs in its design is very recent. This means that it can still have some flaws in relation to recent research in the area of oral proficiency, thus one more reason to use this instrument in scientific research and explore its possibilities, as well as positive and negative aspects.

3. METHOD

In this section, the participants and the instruments 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 adequacy of the new CEFR phonological scale as a tool to measure speech intelligibility. The specific objectives are (a) to investigate how the intelligibility and the comprehensibility constructs are incorporated in the CEFR phonological scale, (b) to compare the performance of experienced raters when using measures of intelligibility and comprehensibility and the CEFR scale, and (c) to correlate the CEFR phonological control scale with intelligibility and comprehensibility measures commonly used in L2 speech research.

The study has the following research questions:

1. How do raters' judgements of L2 learners' intelligibility and comprehensibility relate to the CEFR scale for phonological control?
2. To what extent is the CEFR scale for phonological control helpful in the assessment of intelligibility and comprehensibility by experienced L2 teacher raters?

Furthermore, the study has the following hypothesis:

1. There is a positive, significant correlation between intelligibility measures, comprehensibility ratings and the phonological control scores from the CEFR scale, suggesting that the new descriptors for phonological control adequately reflect the results of intelligibility and comprehensibility assessment for each proficiency level (A1-C2)

3.1 Participants

As shown in Table 2 and 3, the participants of this study were divided into speakers and listeners. Speakers were 16 Brazilian learners of English as an L2, undergraduate students of different majors, with different levels of proficiency. Since the CEFR scale are often used as an instrument to assess proficiency, it is interesting to explore the suitability of this instrument by having speakers with different levels in order for the raters to experience using all the descriptors. Speakers' proficiency was estimated through the Oxford Placement Test, so that we could make sure that we had participants from varied proficiency levels. Speakers also answered a questionnaire to report on their language experiences and uses. As

for recruitment, speakers were invited to participate in this research through social media, contact with teachers and schools from the region and from the Extracurricular language courses at UFSC as well. Their main task was to orally describe an image in English with as much detail as they could. Their description was audio-recorded and the speech samples were assessed by the listeners. Table 2 displays the information about the speakers, including name, university program, time using English throughout the day and their Oxford Placement Test score.

Table 2

Participants: speakers

	Age	Undergraduate/Graduate Programs	Time learning English	Time using English throughout the day	Oxford Placement Test Score
S1	22	Fashion	10 years	2-6 hours	B2
S2	22	Animation	12 years	2- hours	B2
S3	19	English	11 years	2-6 hours	B2
S4	23	English Graduate Program	6 years	6-10 hours	B1
S5	23	Organizational Communication	10 years	10+hours	B1
S6	20	Graphic Design	13 years	2-6 hours	B1
S7	19	Architecture	10 years	2-hours	B1
S8	22	English	15 years	2-6 hours	C1
S9	26	English Graduate Program	6 years	2-6 hours	B1
S10	21	Computer Science	13 years	10+hours	B1
S11	27	English	20 years	2-6 hours	C1
S12	22	Graphic Design	12 years	2-6 hours	B1
S13	24	Medicine	10 years	2- hours	B1
S14	22	Graphic Design	10 years	2- hours	B2

S15	19	Information Systems	14 years	6-10 hours	B1
S16	19	Psychology	5 years	2- hours	B2

Source: the author

As displayed in Table 3, the listeners of this study were 14 English as an L2 teachers, in order to have experienced raters who have a background on language assessment. They listened to the speech samples from the speakers describing an image and assessed their level of phonological control in terms of comprehensibility and intelligibility, through an orthographic transcription task and by using rating scales, as well as assigned them an oral proficiency level (A1-C2) based on the phonological control scale provided by the CEFR. Expert raters (linguists, phoneticians and L2 teachers) have been chosen to work in language assessment in this area due to their language experience and background in the topics of pronunciation, fluency and intelligibility (ISAACS; THOMSON, 2013). Besides, expert raters are expected to be more familiar with language descriptors and terminology, implying that their proficiency rates will be based on their knowledge and experience, and not due to chance or merely subjective. Following the tradition in previous studies, we recruited experienced English teachers to act as listeners in the present research. 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, and they answered a questionnaire to report their language use and teaching experience.

Table 3

Participants: listeners

	Age	Time teaching English	English Teaching Contexts
R1	22	6 years	Language school, outreach programs, one-on-one classes
R2	31	7 years	Language institutes, bilingual school, undergraduate program
R3	29	4 years	Language courses and private classes
R4	23	6 years	Language school
R5	31	12 years	Private classes, regular school and language schools

R6	26	7 years	Regular school and language school
R7	32	10 years	Language school and university
R8	23	4 years	Regular school and language school
R9	34	11 years	Language school and regular school
R10	35	20 years	Regular school, language school, technical courses
R11	32	14 years	Regular school, language school and university
R12	32	14 years	Regular school, bilingual school, language school and private lessons
R13	25	3 years	Language courses and private lessons
R14	32	11 years	Language courses, regular school and technical courses

Source: the author

3.2 Instruments and Materials

The instruments and materials of this research will be detailed in this section. Consent forms, background questionnaires, speech elicitation and listening tasks, and the proficiency test, and the CEFR scale for phonological control (2018 version) will be described and explained throughout the text.

3.2.1 Background Questionnaires

The participants of this study, after signing the consent form, answered a background questionnaire (see Appendix A). For the speakers, there were questions related to their language learning (time and context) and also to their daily language use. They were asked for how long they have been studying English and in which contexts they have studied it (middle or high school, language schools, university, abroad etc.). Moreover, they were asked if they plan to or have the desire to study or work abroad; some parts of this questionnaire can be seen in Figure 1. These questions were designed with the objective of understanding the participants' experiences with the second language as well as to ensure they were potential candidates for taking standardized language proficiency tests using the scales and descriptors similar to the CEFR instrument being investigated in this research.

For the listeners, the questions addressed their time teaching English and the environments they have worked in, as well as their daily language use (see Appendix A). The listeners answered this questionnaire to guarantee that they are L2 teachers and for us to

understand in what contexts they work. The questions were created to understand their experience in the area.

3.2.2 Consent Forms

The project was submitted to the Ethics Board² and received the approval to start the data collection. There are two different consent forms for the participants of this research: one for the speakers and one for the listeners. The consent forms explained the objectives of the study, the design and the instructions, as well as stated that the participants could leave the research at any time. The forms also provided the researchers' contact information and stated the possible risks and benefits from participating in the study.

The speakers' consent form explained all the steps, from the background questionnaire until the speech recording session, stating that their image would not be used in the study. The listeners' consent form provided instructions for all the assessment tasks (intelligibility, comprehensibility and phonological control rating).

3.3.3 Proficiency Test

In order to estimate the speakers' proficiency before assessing their intelligibility, an adapted version of the Oxford Placement Test was administered online through a Google form, as can be seen in Figure 3. The paper and pen 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. (Appendix B). The speakers performed the placement test while in a Zoom meeting with the researcher, with their cameras turned on. In the same meeting, after finishing the placement test, they went through the speech elicitation task.

The researcher examined the answers of each participant with the answer sheet and assigned the proper proficiency level according to their scores. Table 2 displays the proficiency level of each speaker, in line with the scoring from the Oxford Placement Test. All of the speakers received either a B1 level, B2 or C1, for their reading, vocabulary and grammar L2 proficiency.

² CAAE: 48418621.9.0000.0121. Parecer: 5.001.578

Questions 1-5.
Where can you see these notices?
For questions 1 to 5, mark one letter A, B or C.

1 "You can look, but don't touch the pictures"

in an office

in a cinema

in a museum

2 "Please give the right money to the driver"

in a bank

on a bus

in a cinema

3 "No parking please"

in a street

on a book

on a table

Figure 1: Oxford Placement Test online version extract

Source: Adapted from Allan (2004)

3.3.4 Speech Elicitation

The first task was an Image Description test to elicitate the speech of the participants. The participants saw an image of a working space (Figure 4) 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 elicit speech (DERWING; MUNRO; THOMSON, 2008; ISAACS; TROFIMOVICH, 2012; SILVEIRA; MARTINS, 2020). Since the raters were expected to transcribe the 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.



Figure 2: Speech Elicitation Image

Source: the author

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 motifs, topics of conversation happening, possible relationships between the characters and so on. It is a picture that can be used to elicit a range of language performance because it can be described using either simple vocabulary and language structures or more complex language.

3.3.5 Listening Tasks

The raters listened to the speech samples and assessed them in terms of intelligibility (orthographic transcription of utterances) and comprehensibility (rating scale). The intelligibility score consisted of the percentage of words correctly transcribed by the listeners, and the comprehensibility rating consisted of listeners' perception of difficulty of understanding the utterances. Finally, the listeners used the CEFR descriptors to rate

phonological control, which is part of the speaking proficiency scales (CEFR, 2018). These three steps were on the same Google form, in different sections (Appendix C). 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 (Figure 3). 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 and, right after this chart, a multiple-choice item from which the listeners can select the appropriate level of the phonological scale. These steps were repeated for each speaker.

Participant 1

(audio sample)
(image being described)

Transcribe every word you hear from the speech sample. *

A sua resposta

How difficult was it to understand the speech? *

1 2 3 4 5 6 7 8 9

no difficulties extremely difficult

Read the phonological control scale and assign one level for the participant *

PHONOLOGICAL CONTROL		
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 teacher message are clear and precise. Intelligently and effectively compensate for and minimisation of meaning are not affected in any way by features of accent that may be retained from other language(s).	Can articulate virtually all of 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 of the sounds of the target language, some features of accent retained from other language(s) may be noticeable, but they do not affect intelligibility.	Can articulate virtually all of the sounds of the target language with a high degree of control. He/she can usually self-correct if he/she noticeably mispronounces 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 he/she means to say.
B2 Can generally use appropriate intonation, place stress correctly and articulate individual sounds clearly, accent tends to be influenced by other language(s) he/she speaks, 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 teacher 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 he/she intends to convey, though with some influence from other language(s) he/she speaks.
B1 Pronunciation is generally intelligible, can approximate intonation and stress at both sentence and word levels. However, accent is usually influenced by other language(s) he/she speaks.	Is generally intelligible throughout, despite regular mispronunciation of individual sounds and words he/she is less familiar with.	Can convey teacher message in an intelligible way in spite of a strong influence on stress, intonation and/or rhythm from other language(s) he/she speaks.
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 other language(s) he/she speaks 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 other language(s) he/she speaks. 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 concerned. Can reproduce correctly a limited range of sounds as well as the stress on 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 giving 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 other language(s) he/she speaks. Teacher interlocutor needs to be collaborative.

Page 136 • CEFR Companion Volume with New Descriptors

You can access the image for better reading here <https://bit.ly/3mYQGo7>

- A1
- A2
- B1
- B2
- C1
- C2

Figure 3: Listening Tasks

Source: the author

The final section of the Google form was the reflective questionnaire about the listeners' experience using the phonological scale, as can be seen in Figure 4. This questionnaire had open and closed questions in which they were instructed to be reflective about and to report their experiences with as much detail as they could, in order to provide data regarding the use of the descriptors to assess L2 speakers.

Report your experience using the CEFR's Phonological Control Scale

How difficult it was to use the scales?

1 2 3 4 5 6 7 8 9

no difficulties extremely difficult

Do you agree with the descriptors for each level?

Yes

No

Do you think the descriptors are too detailed?

Yes

No

Do you think the descriptors lack in detail?

Yes

Figure 4: Sample of the questionnaire to report the experience using the phonological control scale

Source: the author

3.4 Procedures For Data Collection

After having the project approved by the ethics board, the participants were recruited - through social media, contact with teachers and schools from the region - and those who accepted participating in the study received an email with the consent form, the background questionnaire and instructions to schedule an online meeting through Zoom, which allows high quality image and audio recordings, to go through the proficiency test and the speech elicitation task. During the meeting session, the speakers received a link to a Google form for the Oxford Placement Test, which they answered during the meeting. After finishing the proficiency test, the researcher shared a slide presentation with the instructions for the image description test. The speaker had a maximum of 30 seconds to plan the description, without taking notes. The speaker was instructed to describe the image, and the complete session was recorded. The participants' were informed that their image was not to be used at any time, only their speech data would be the object of analysis. The participants needed around 60 minutes to complete all the tasks.

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

The listening tasks were also done entirely online, through a Google Form containing the speech samples and sections to assess intelligibility and comprehensibility, as well as with a space for open and closed questions regarding the use of the CEFR phonological control scale. As demonstrated in a study by Silveira and Martins (2020), Google Forms can be used as an instrument for assessment of speech samples by listeners, using both holistic and analytic criteria to rate oral proficiency.

In the present study, the listeners were recruited - through social media and contact with teachers and schools as well - and after having accepted the invitation to contribute with the study, they received an email with instructions and a link to access the Google Form, which contained three sections: (a) consent form, (b) background questionnaire, and (c) listening tasks (intelligibility task, comprehensibility task, and phonological control

assessment). Listeners were expected to complete all tasks in 90 minutes, but they could work at their own pace.

3.4.1 Intelligibility Assessment

For the intelligibility assessment, the raters were asked to wear headphones and orthographically 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. The researcher transcribed every sample, in order to compare to the raters' transcription and be able to assign intelligibility scores. All the transcriptions can be seen in Appendix D. The percentage of correct words from each transcription corresponds to the intelligibility results, following Derwing and Munro (2005). Listeners were able to listen to the speech samples as many times as they found necessary as they were performing the transcription.

3.4.2 Comprehensibility Assessment

For the comprehensibility measurement, there was a Likert scale for the raters to decide how difficult it was to understand the utterances, also suggested by Derwing and Munro (2005). In the Likert scale ranging from 1 to 9, "1" referred to "no difficulties" and "9" referred to "extremely difficult". 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. Rating scales to assess comprehensibility are quite common in the field. Saito, Trofimovich and Isaacs (2015) used scalar judgements to assess comprehensibility; the authors used a 9-point Likert scale where 1 referred to "very easy to understand" and 9 referred to "very hard to understand". The listeners were oriented to use the entire scale, and the researchers stated that the scale represented a range of ability levels, from natively like speakers to beginners.

Despite the comprehensibility scale labels sometimes differing in the literature, where "1" sometimes refers to easy or difficult, they all work the same in the sense of describing the ease or the difficulty in comprehending foreign speech. In this study, nine-point Likert scales were chosen as an appropriate method to measure comprehensibility, since it has been established in the literature as a successful and reliable tool.

3.4.3 Assigning oral proficiency levels with the phonological control scale

After completing the intelligibility and comprehensibility tasks, the raters were asked to assign proficiency levels to the speakers, using the CEFR descriptor scale for phonological control (see Appendix E), where A1 is described as “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 concerned. Can reproduce correctly a limited range of sounds as well as the stress on simple, familiar words and phrases.” and C2 as “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 his/her message are clear and precise. Intelligibility and effective conveyance of and enhancement of meaning are not affected in any way by features of accent that may be retained from other language(s)”. (COUNCIL OF EUROPE, 2018, p. 136).

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 section in the form. The listeners were oriented to reflect about their choices while using the scale, as well as to pay close attention to the descriptors in order to make a conscious choice regarding the speakers’ proficiency level. In order for their answers on the questionnaire to be reliable, it was necessary for them to reflect while in the process of reading the scale and deciding on the best option for each speaker. After assigning the phonological control levels to the speakers, the raters reported on their experience using the CEFR phonological control scale by answering open and closed questions, to assess usage and feasibility of this tool. The questions were related to the difficulty of using the descriptors, the content of the descriptors (if they thought they were too detailed or not detailed enough), if they agreed with the statements in the descriptors, if they thought the scale was a good source to use in the assessment of pronunciation, and if and how they thought the scale could be improved. There was also a space for them to comment about anything related to the use of the descriptors.

3.5 Procedures for Data Analysis

Considering the nature of the two research questions (i) How do raters' judgements of L2 learners’ intelligibility and comprehensibility relate to the new CEFR scale for phonological control? (ii) To what extent is the improved CEFR scale helpful in the

assessment of intelligibility and comprehensibility by experienced L2 teacher raters? The most appropriate way of analyzing the data is to look at the relationship between variables, and to examine raters' responses to the closed and open questions about their experience with the CEFR phonological control scale. The quantitative variables for the correlational analysis, in this case, are (i) intelligibility scores, (ii) comprehensibility scores, (iii) scores for phonological control using the CEFR scale. Pearson Correlation was used to explore the relationship between variables, mentioned by Pallant (2015) as the adequate statistical test for this purpose. 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. This type of test generates a Cronbach's alpha coefficient, which is also mentioned by Pallant (2015) as an appropriate indicator of consistency across raters.

Qualitative analysis was also conducted by referring to questionnaire data that can help interpreting the results, and special attention was given to the raters' views on the usefulness of the CEFR phonological control scale, searching for response patterns unveiled by the responses to the open and closed questions.

4. RESULTS AND DISCUSSION

4.1 Introduction

The objective of this study is to investigate the relationship between intelligibility, comprehensibility and the CEFR phonological scale as a tool to assess these features of oral proficiency, in order to see if this instrument is useful for assessment. The research questions are: a) How do raters' judgements of L2 learners' intelligibility and comprehensibility relate to the CEFR scale for phonological control?, and b) To what extent is the CEFR phonological scale helpful in the assessment of intelligibility and comprehensibility by experienced L2 teacher raters?

To try to answer these questions, we collected data from 16 speakers of English as a second language; each one of them described the same image in English and their speech samples were recorded. These samples were then assessed, in terms of intelligibility and comprehensibility, by 14 experienced raters, teachers of English as a second language. The raters transcribed the speech samples and an intelligibility score was created by counting the amount of words transcribed correctly. They also assigned a score on a Likert Scale for comprehensibility (1= no difficulties to understand, 9= extremely difficult to understand) and, after that, assigned a CEFR Phonological Scale level for each of the participants (A1, A2, B1, B2, C1, C2). The raters were provided with a table containing the descriptors for the evaluation of phonological control.

To analyze and interpret the results of the collected data, we first ran an interrater 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 chapter, starting with Cronbach's interrater reliability analysis.

In order to answer the research questions, we are going to present both the quantitative and qualitative results of this research altogether with a discussion. This chapter is going to be organized into two major sections. Section 4.2 attempts to answer the first research question and focuses on the quantitative results, starting with the interrater reliability results and then moving to the descriptive statistics. Afterwards, we are going to present the correlation analysis between the three variables. Finally, we are going to discuss the results

per speaker and per listener, investigating the consistencies, discrepancies and other relevant information that arose from the tests. In section 4.3, we attempt to answer the second research question by examining the survey data and exploring the quantitative data provided by the listeners after using the CEFR scale.

4.2 Correlational Analysis Results

This section will present the quantitative results of this study such as the interrater reliability test, correlational analysis, descriptive statistics and individual results, alongside with tables and figures to illustrate the study and help with the understanding of the statistical results. Our analysis is guided by the question: How do raters' judgements of L2 learners' intelligibility and comprehensibility relate to the CEFR scale for phonological control?

4.2.1 Interrater Reliability

Considering the data of this study was provided by raters, it is necessary to examine if they agree with each other and to what extent (LARSON-HALL, 2010). We conducted an interrater 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 interrater reliability is called Cronbach Alpha, which is a measurement of intraclass correlation. According to Cortina (1994, apud LARSON-HALL, 2010), this coefficient estimates the internal consistency between subjects and items; in this study we are measuring the consistency between raters.

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 Appendix F.

4.2.2 Descriptive Statistics

Table 4 displays the descriptive statistics for this study. For the intelligibility score, the percentage for each participant was calculated by the amount of correct words transcribed

by the raters. As the results displayed in Table 4 show, the minimum score for the intelligibility variable was 94,71 and the maximum 99,86, meaning that all participants were highly intelligible. For the second variable, comprehensibility, the raters used a scale in which the range varied from 1 (no difficulties to understand the speech) to 9 (extremely difficult). The minimum score assigned by the raters was 1 and the maximum 4,07; this means that the raters assigned harsh comprehensibility rates even though they could understand the speech fairly well, as demonstrated by the high intelligibility scores.

Table 4

Overall scores for intelligibility, comprehensibility, and CEFR Phonological Scale (combined means for the 14 listeners).

	N	Minimum	Maximum	Mean	Standard Deviation
Intelligibility	16	94,71	99,86	97,9643	1,29730
Comprehensibility	16	1,00	4,07	2,2723	,77073
Phonological Control	16	2,43	4,93	3,4688	,68758

Source:the author

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 ("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.") and 6 to 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)."). According to Table 4 the minimum score assigned for phonological control was 2,43 and the maximum 4,93, which demonstrates that there was an average of B1 levels ("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.”) being assigned to the speakers. In other words, most speakers were judged as possessing a pre-intermediate proficiency level regarding phonological control.

As table 4 displays, the standard deviation for the intelligibility variable was 1.29; for comprehensibility, .77 and, for phonological control, .68. The standard deviation for the three variables was small; this means that there was not a wide range of scores and rates for the speakers, as they were close to the mean value. Figures 5, 6 and 7 bring a visual representation of the results for the three variables. These boxplots reveal the small range of variance for the three variables, as shown by the symmetrical boxes and the short whiskers in each figure. There is only one outlier identified in the intelligibility variable (Figure 5): speaker 4, who received the lowest intelligibility average score (94.71%). The scores assigned to this speaker will later on be investigated and discussed in this chapter.

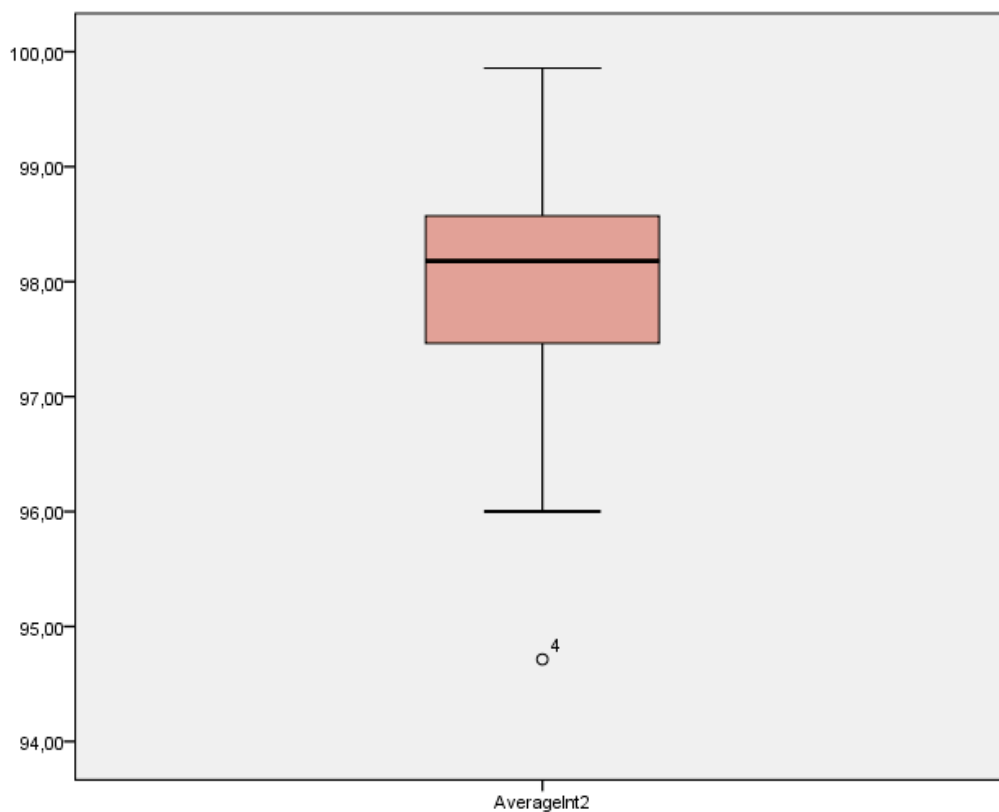


Figure 5: Standard deviation for the intelligibility variable

Source: the author

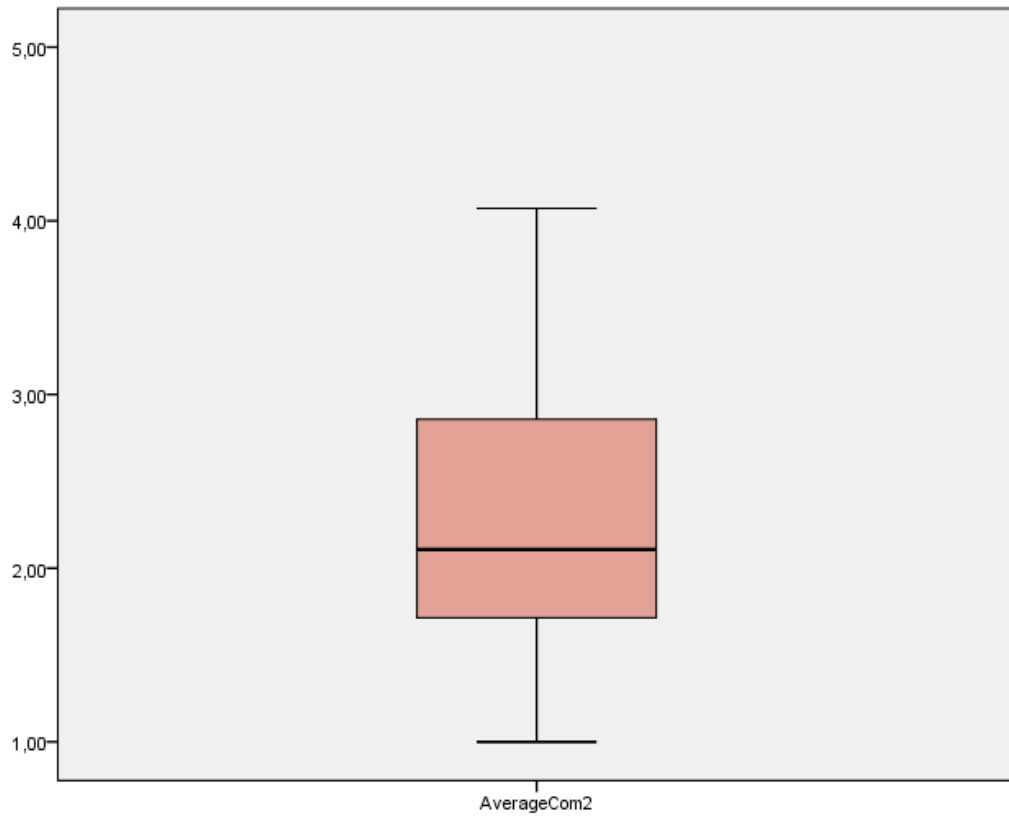


Figure 6: Standard deviation for the comprehensibility variable

Source: the author

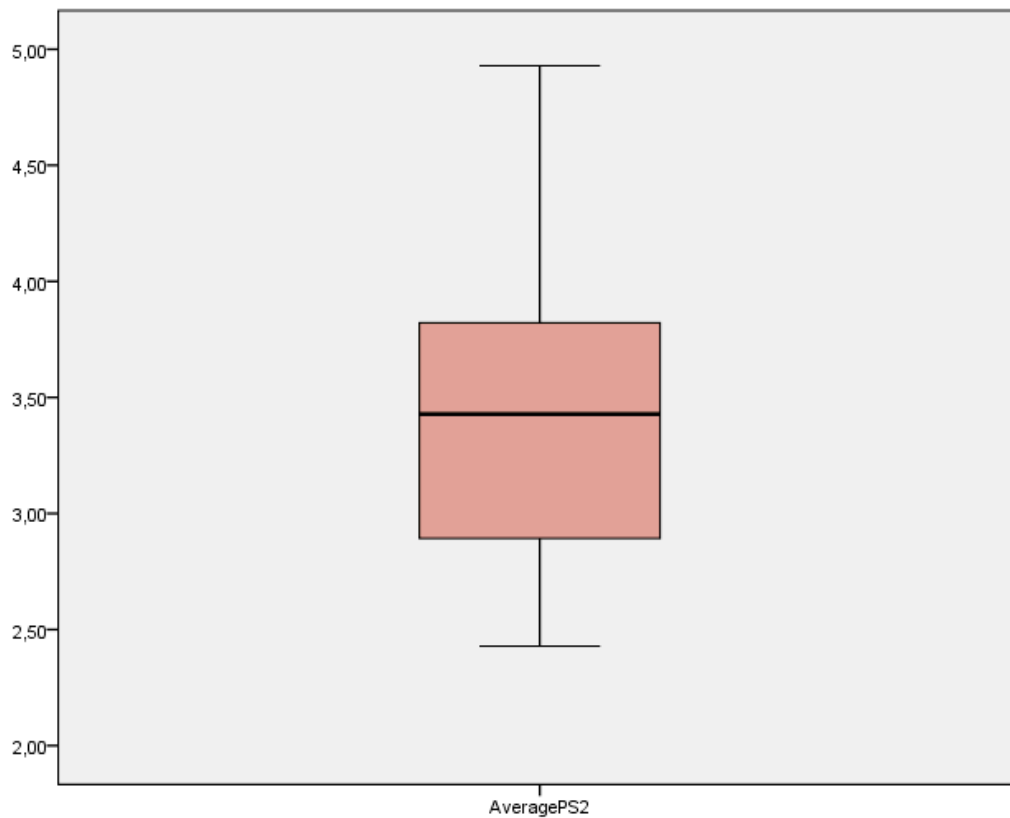


Figure 7: Standard deviation for the phonological control variable

Source: the author

4.2.3 Correlational Analysis

In order to answer RQ1 (How do raters' judgements of L2 learners' intelligibility and comprehensibility relate to the CEFR scale for phonological control?), we ran Pearson correlation using the three variables, namely, intelligibility, comprehensibility, and phonological control. As Table 5 and Figure 8 show, there is a highly significant relationship between the average intelligibility and the average comprehensibility variables, $r(14) = -.80$, $p = <.001$. This means that the more intelligible the speaker is, the easier it is to comprehend him/her. The negative sign in this correlation is due to the 9-point Likert Scale used for the Comprehensibility measures, where 1 means no difficulties to understand and 9 means extremely difficult to understand.

Table 5

Pearson Correlation

	Average Comprehensibility	Average Phonological Control
Average Intelligibility	-,806**	,428
	,000	,098
	16	16
Average Comprehensibility	-	-,699**
		,003
		16

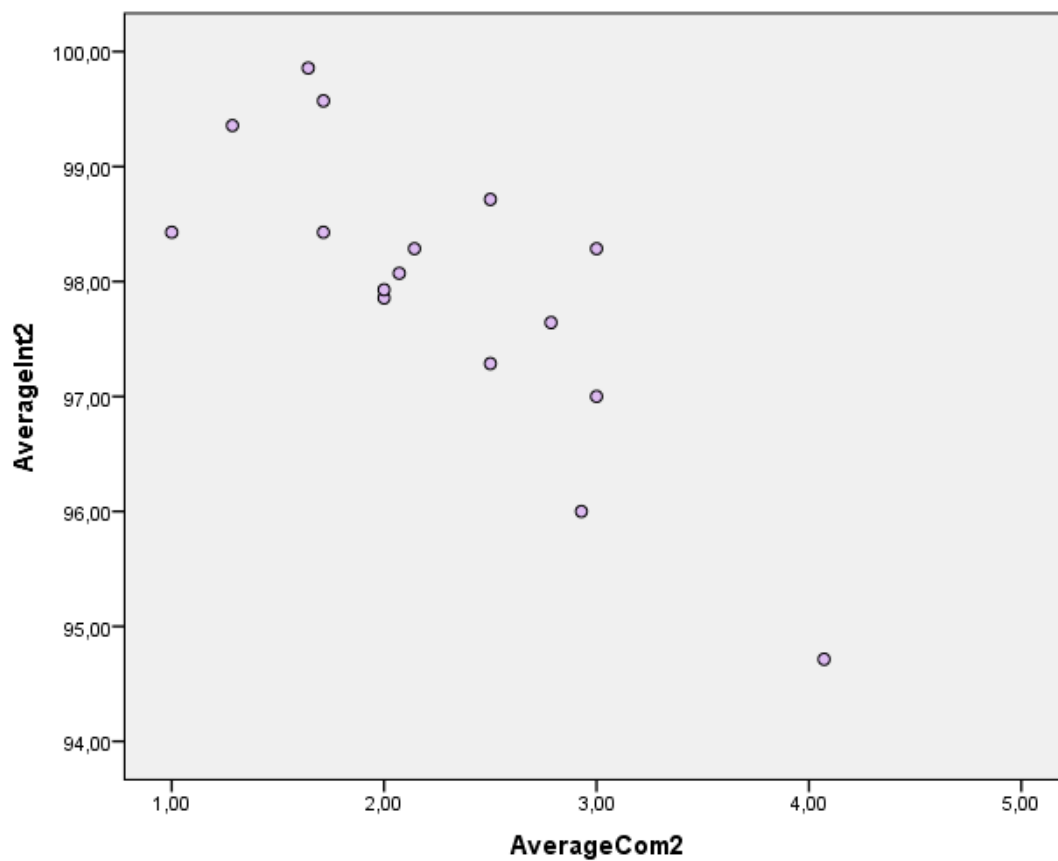


Figure 8: Correlation between intelligibility and comprehensibility

Source: The author

Table 5 and Figure 9 also shows that there is a strong, significant relationship between the average comprehensibility rates and the average phonological control scores, $r(14) = -.69$, $p = .03$. This means that the easier to comprehend the speaker is, the higher was the phonological control score they assigned by the raters using the CEFR phonological scale. Again, the negative correlation is a consequence of the 9-point Likert Scale where 1 means no difficulties to understand and 9 means extremely difficult.

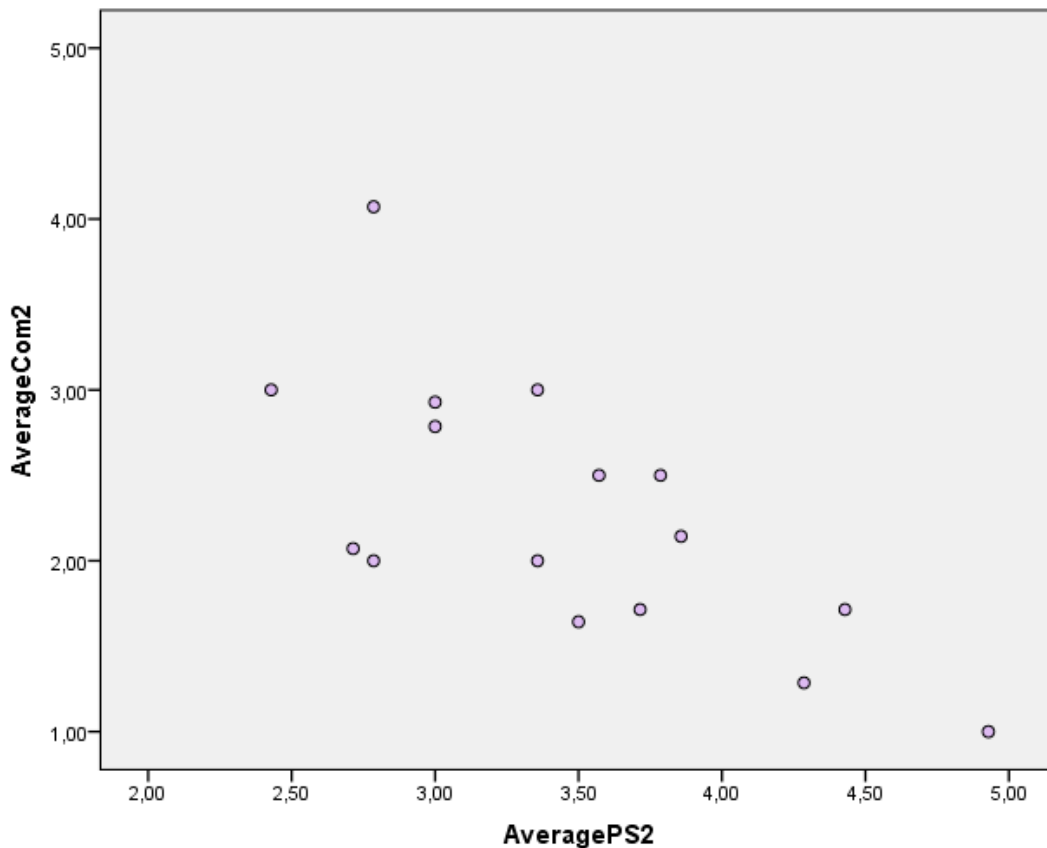


Figure 9: Correlation between comprehensibility and phonological control

Source: The author

Finally, as demonstrated in Table 5 and Figure 10, there is a moderate positive relationship between the average intelligibility score and the average phonological control score, $r(14) = .42$, $p = .09$. This means that the more intelligible speakers received higher ratings from the listeners as regards phonological control. However, no statistical significance was obtained, meaning that the comprehensibility variable is a better indicator of phonological control than intelligibility.

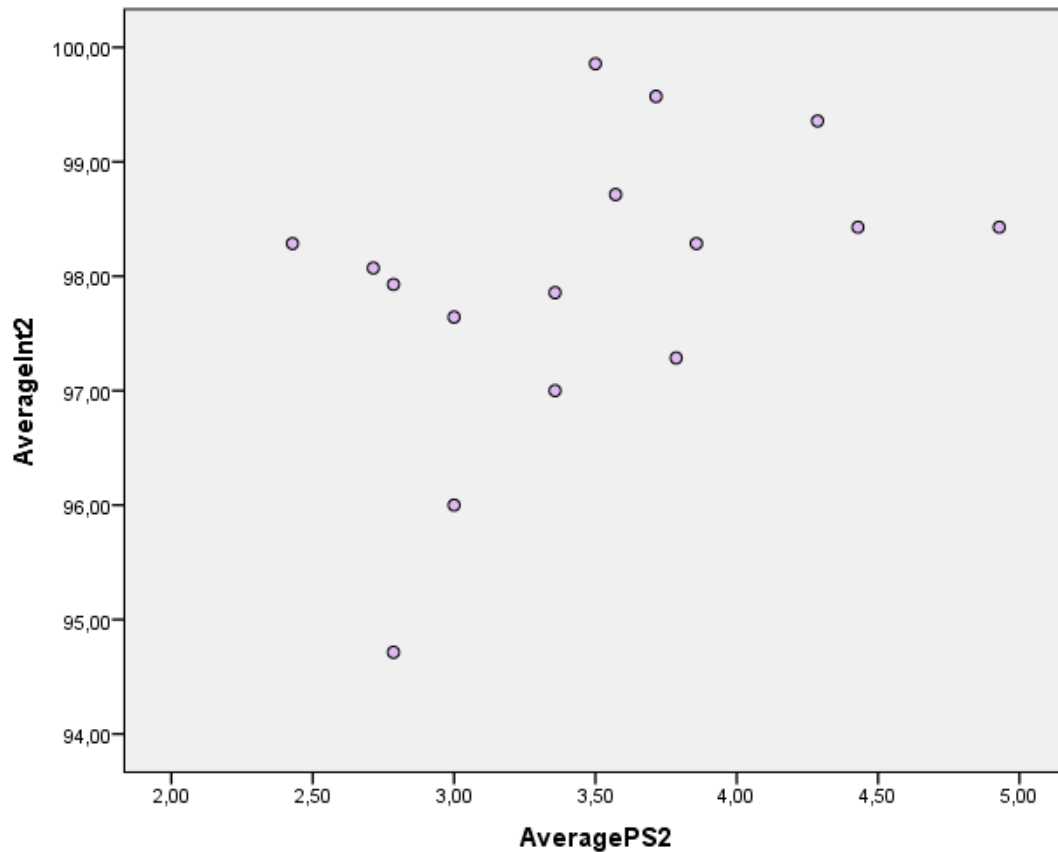


Figure 10: Correlation between intelligibility and phonological control

Source: The author

Hypothesis 1 stated that there would be a positive, significant correlation between intelligibility measures, comprehensibility ratings and the phonological control scores from the CEFR scale, suggesting that the descriptors for phonological control yield results similar to those provided by alternative measures of speech intelligibility and comprehensibility commonly used in L2 speech research, thus allowing adequate assessment of L2 pronunciation levels (A1-C2). Hypothesis 1 is partially confirmed by the correlational analysis results. Regarding the comprehensibility and phonological control variables, and the intelligibility and comprehensibility variables, significant correlations were found. As can be seen in Table 5, the only correlation that did not show a significant relationship was between the intelligibility variable and the phonological control variable, implying that intelligibility was not a reliable indicator of phonological control for the participants of this study. This is probably due to the fact that all speakers received high intelligibility scores and there was not a range of intelligibility scores to discriminate across proficiency levels.

4.2.4 Individual results

In order to understand the way the listeners evaluated the speakers' performance, it is important to examine the individual rates and scores assigned for each variable, as displayed in Tables 6, 7 and 8. This subsection will focus on the individual results for speakers and raters, highlighting consistencies and discrepancies amongst the assessments.

Table 6

Scores assigned for intelligibility (percentage of words transcribed correctly)

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14
S1	100	100	100	98	98	100	98	100	97	100	96	97	98	96
S2	96	94	100	100	100	98	98	100	100	96	94	94	100	100
S3	100	96	96	100	98	98	96	100	98	98	98	100	100	100
S4	94	97	94	97	91	86	88	97	100	97	97	94	94	100
S5	100	94	100	97	94	94	94	100	89	94	100	94	100	94
S6	100	100	100	100	100	100	100	100	97	100	97	100	100	100
S7	100	96	98	98	100	96	93	100	96	98	98	96	98	100
S8	98	100	100	98	100	100	98	100	100	100	97	100	100	100
S9	100	93	100	95	97	97	95	97	100	97	100	100	100	100
S10	98	100	96	98	96	100	98	96	98	93	93	96	100	96
S11	98	93	100	97	94	98	96	96	100	98	98	98	98	98
S12	97	100	100	97	97	100	100	97	97	100	94	97	100	97
S13	100	100	100	100	100	100	100	100	100	100	98	100	100	100
S14	100	100	100	94	98	98	98	100	98	98	98	100	100	100
S15	97	100	100	97	100	100	100	100	97	97	97	97	97	97
S16	98	100	98	98	98	96	96	100	100	98	94	100	100	100

R: rater; S: speaker

Source: the author

Raters' consistency can be observed in the scores; speaker 6 and speaker 13, for example, were one hundred percent correctly transcribed by almost all of the listeners. Overall, all speakers were highly intelligible, according to table 6; the lowest percentage being 93 for speakers 9, 10 and 11, which is still a very high percentage. The highly significant correlation between intelligibility and comprehensibility measures demonstrate that the more intelligible a speaker is, the easier it is to comprehend their speech. Table 7

shows the scores for comprehensibility, where number 1 refers to no difficulties to comprehend, and number 9 extremely difficult.

A different situation was observed for the comprehensibility variable, as Table 7 shows. It is interesting to notice that some raters did not consider some speakers comprehensible while the intelligibility percentage was very high. For example, raters 1 and 3 were able to understand 94% of the speech sample provided by speaker 4; however, they both assigned number 5 for the comprehensibility measure. For the same speaker, rater 4 understood 97% of the speech and assigned a comprehensibility score of 5 as well. One possible reason for that is that they could have transcribed almost every word correctly because they listened to the sample more than once; and, having to listen to the audio twice or more might have made them consider that the speech was not that easy to comprehend.

Table 7

Comprehensibility scores assigned by the raters

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14
S1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
S2	1	3	1	2	5	2	2	3	1	1	3	2	1	1
S3	1	1	3	2	1	3	3	1	1	1	2	1	2	2
S4	5	2	5	5	8	5	3	5	2	2	5	3	5	2
S5	1	5	1	4	6	5	3	3	1	2	3	3	2	2
S6	2	1	1	2	1	1	3	1	3	2	4	1	1	1
S7	3	3	2	2	6	2	4	3	3	2	5	1	2	1
S8	1	1	1	1	1	1	3	1	1	1	3	1	1	1
S9	2	1	1	4	2	1	3	2	1	2	4	2	2	1
S10	3	5	2	3	2	4	2	2	5	3	5	3	2	1
S11	1	6	1	4	2	2	3	4	2	1	3	2	2	2
S12	1	6	1	3	1	1	3	1	1	1	5	2	2	1
S13	2	1	1	3	1	2	2	1	1	2	3	1	2	1
S14	1	4	1	3	2	2	3	1	5	4	4	1	3	1
S15	1	5	3	3	1	2	3	3	2	5	6	2	3	3
S16	4	1	1	3	2	1	3	1	2	2	5	1	3	1

Source: the author

The results regarding the phonological control variable demonstrate that the intelligibility measure is not a good predictor of phonological control as measured with the

CEFR scale. According to table 8 some raters assigned beginner levels (A1 and A2) for speakers that were considered highly intelligible. For instance, raters 4, 9 and 12 designated level A2 for speaker 4, but their intelligibility percentage, respectively, was 98, 98 and 96 for speaker 4. Another example is regarding speaker 15, who was assigned level A1 by raters 8 and 10; these raters displayed intelligibility scores of 100% and 97%, respectively, for speaker 15. This means that the raters did not consider high intelligibility as a feature that represents higher levels of phonological control; a person can be vastly intelligible but lacking control of other features of speech (e.g., speech rate, segmental accuracy fluency, intonation) in order to be considered an intermediate or advanced second language speaker in terms of phonological control.

The opposite can also be inferred: a speaker from a beginner level, with a limited repertoire, can sustain a conversation in terms of intelligibility; their interlocutor will be able to understand their speech and communication will be established. These results are in-line with Derwing and Munro (1995), when they mention that a speaker can be intelligible despite other features of speech, such as accent, in their case.

Table 8

Phonological Scale CEFR levels assigned by the raters

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14
S1	5	5	6	5	5	4	5	5	5	5	5	5	4	5
S2	4	3	5	4	3	2	4	3	2	4	4	3	3	3
S3	5	6	3	4	6	3	4	5	5	5	4	5	4	3
S4	4	4	2	3	1	2	4	2	2	3	3	3	3	3
S5	4	3	6	3	2	1	4	1	3	2	4	3	3	3
S6	4	4	5	4	5	4	4	2	2	2	3	5	4	4
S7	4	3	4	3	3	3	3	2	2	1	3	5	3	3
S8	4	5	1	4	5	4	4	5	5	5	3	5	5	5
S9	3	3	4	2	4	3	4	2	3	1	3	3	2	2
S10	4	3	4	2	5	3	4	4	2	4	3	2	4	3
S11	4	3	5	3	5	4	4	3	3	4	4	4	4	3
S12	4	3	5	2	5	2	4	2	2	1	3	2	2	1
S13	4	5	4	3	5	2	5	3	3	1	4	5	3	2
S14	4	4	5	3	5	3	4	5	2	1	3	5	3	3
S15	4	2	3	2	5	2	4	1	2	1	2	2	2	2
S16	4	5	5	3	5	4	4	5	2	3	3	5	3	3

Source: the author

Notes: 1=A1, 2=A2, 3=B1, 4=B2, 5=C1, 6=C2

Comprehensibility appears to be a better predictor for the phonological control levels assignment. Table 7 demonstrates that speaker 1 received a “no difficulties to understand” score from all of the raters. Consistently, the majority of them assigned a C1 level for this speaker, meaning that the highly comprehensible speech was taken into consideration as an important feature when designating a CEFR level for phonological control. The same can be observed with speaker 8, who generally received number 1 scores for comprehensibility (no difficulties to understand) and was later assigned with mostly C1 levels. This strong relationship between comprehensibility and the use of the phonological scale can be noticed regarding beginner levels as well; speaker 4 received a number 5 score for comprehensibility from many raters and later obtained mostly A2 and B1 levels from the CEFR phonological scale. Trofimovich and Isaacs (2012) concluded that grammatical and lexical aspects of language are related to comprehensibility and thus influence the perspective listeners have on speakers’ oral production; although in the present study the specific aspects that permeate second language speech were not investigated, grammar and vocabulary could have played a role during the raters’ analysis.

There are some disagreements among raters when assigning a phonological control level for the speakers. Speakers 5 and 7, for instance, has a range going from 1 to 6, as table 8 shows. For speaker 15, raters assigned scores from 1 to 5 and, and for speaker 16, from 2 to 5. Rater 4 sometimes demonstrates a discrepancy from the other raters; when assigning a CEFR level for speaker 8, for instance, the majority of raters decide on B2 and C1 levels, while rater 8 designates an A1 level for this participant. A similar situation can be seen regarding speaker 5, who generally received A1, A2 and B1 levels from the other raters, but obtained a C2 level from rater 4.

Rater 11 appears to be the harshest evaluator, according to table 7. This rater mainly attributed 4 and 5 scores for comprehensibility; while most peers considered “no difficulties” to understand the speakers. Regarding the CEFR levels for phonological control, speaker 8 was assigned mostly B2 and C1 levels, however, rater 11 designated a B1 level for this participant, according to table 8.

4.2.3 Summary

The results and discussion for the quantitative analysis of this research was presented in this section. It was demonstrated that there was a highly significant correlation between the variables intelligibility and comprehensibility, and between the variables comprehensibility and phonological control. A moderate correlation was seen between the variables intelligibility and phonological control, which demonstrates that comprehensibility is a better predictor of phonological control scores than intelligibility, and it is considered a relevant feature for raters when assigning a phonological control level for speakers.

The first hypothesis, which stated that there is a positive, significant correlation between the intelligibility measure, comprehensibility ratings and the phonological control scores from the CEFR scale was partially confirmed, considering that a significant relationship was not found between the intelligibility scores and the phonological control levels designated to speakers.

The second research question and the second hypothesis will be explored in the next subsection, which will examine the survey data of this research, investigating closed and open questions in a questionnaire provided for the raters. Each question and its answers will be presented and discussed separately, while making inferences with the quantitative analysis and the review of literature that ground this study.

4.3 Survey Results

To further understand the usefulness and applicability of the CEFR phonological control scale to the assessment of intelligibility and comprehensibility of ESL speakers, we developed a questionnaire with open and closed questions for the raters to respond. The complete questionnaire can be seen in Appendix C. The raters' responses to the questionnaire helped us to investigate and answer the second research question of this study: To what extent is the CEFR scale for phonological control helpful in the assessment of intelligibility and comprehensibility by experienced L2 teacher raters? Figures 11, 12, 13 and 14 summarize the responses to the closed questions.

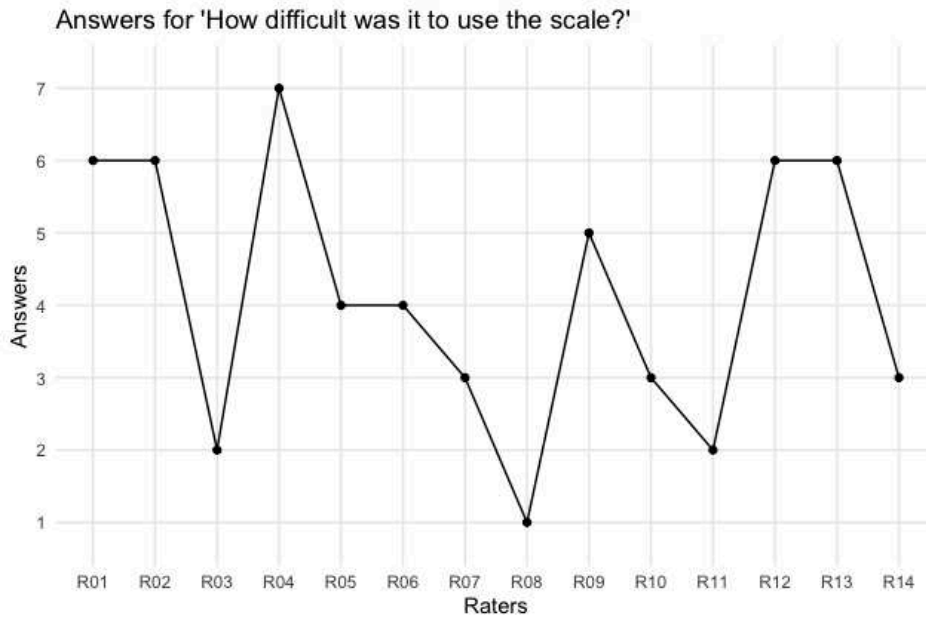


Figure 11: Answers for question 1

Source: The author

The first closed question was “how difficult was it to use the scale?” to which the listeners answered through the same 9-point Likert scale (1 - no difficulties and 9 - extremely difficult). The average rate obtained was 4 (SD = 1,8). Only rater 8 chose the option “no difficulties”. The most chosen rate was number 6, selected by four different raters. Nobody decided on rate 9 (extremely difficult), and the higher rate selected was 7 by one participant (rater 4).

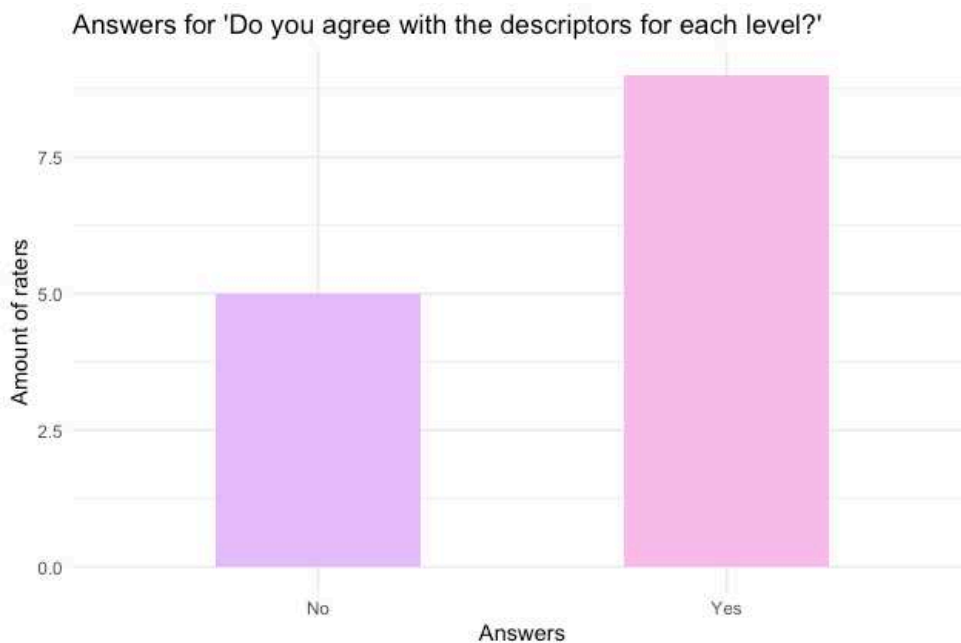
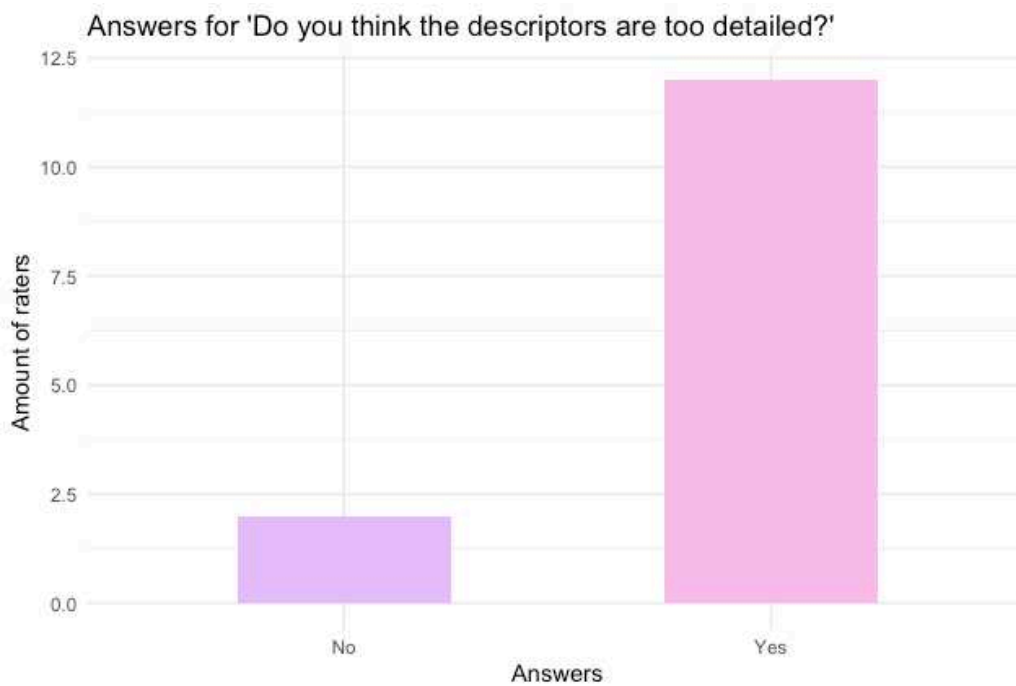


Figure 12: Answers to question 2**Source: The author**

The second closed question aimed to discover if the raters agreed with the descriptors. As shown in Figure 12, nine participants answered “yes” and five answered “no”. Their answers will be justified in the final open questions, where they provided details about their experiences. Twelve participants answered “no” to the third question: “do you think the descriptors are too detailed?” (Figure 13) and eight participants answered “yes” to the fourth question: “do you think the descriptors lack in detail?” (Figure 14), meaning that the majority of them considered the descriptors adequate regarding the explanation of the levels, but still are missing some information they judged to be important for the assessment.

**Figure 13: Answers to question 3****Source: The author**

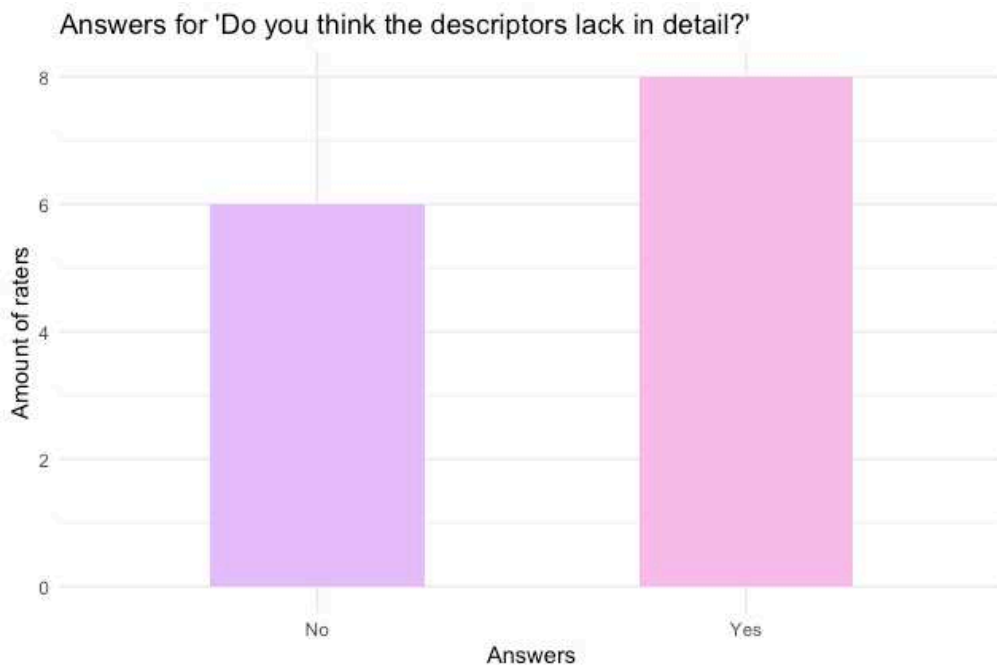


Figure 14: Answers to question 4

Source: The author

The first open question “do you think the phonological control scale is a good source to assess pronunciation? why?” unveiled very detailed opinions from the raters. Rater 8 stated that

“Yes, I believe it is extremely helpful, However, it is important to state, during the evaluation that this scale is only related to phonological control, not grammar, or reading levels. I sometimes caught myself placing a wrong scale based on its grammar, even though I was able to understand every word they've said (all this considering the introductions). As this scale (A1-C2) is also used to measure other language levels, it might be confusing for the teacher and students separate things. But in terms of practicability, it's indeed very good and it works considering that pronunciation is something hard to put on a scale without getting caught on linguistics biases.”

This report is in-line with the results regarding the relationship between intelligibility, comprehensibility and phonological control, and how this relationship between the variables was not significant when it comes to intelligibility and phonological control. This suggests that the raters were taking into consideration (or trying not to) other features of speech, such as grammar. Furthermore, it brings to light a relevant reflection on how difficult it may be

sometimes for teachers and students to understand and separate the definitions present in such assessment scales.

Another interesting response comes from rater 11, who stated that

“It depends. If the goal is assessing their pronunciation to find someone's level in the CEFR, then I think it's appropriate. However, if the goal is assessing intelligibility, then I don't think they help much, because all of the participants I heard were very much intelligible, but they'd likely fall into different levels in CEFR. I think the greatest difference was that they had different levels of control of prosodic features, with many sounding very flat. So they do have "worse" pronunciation in some sense, but they are still intelligible.”

This participant provides one possible explanation for the negative correlation between intelligibility and phonological control. Rater 11 explains that all speakers were highly intelligible, as demonstrated in table 6, where they all had an above 90 percentage of correct words transcribed. However, this rater stresses that “they had different levels of control of prosodic features, with many sounding very flat”, meaning that the raters could understand the words, as it was proposed of them, but they could still notice the differences in the speeches, being them segmental features (pronunciation of vowels and consonants) or suprasegmental/prosodic (rhythm, stress, tone), which was the case for rater 11.

One of the raters yielded a negative response to the same question. Rater 13 shared that

“Not really. It is very vague in some aspects. For example, what does it mean "virtually all sounds of the language"? It is inconsistent in the descriptions of each level. For example, stress and intonation is mentioned in the majority of the levels, but is left out on others. Also, sometimes there are behavioral descriptions of a conversation, like "conversational partners will need to ask for repetition from time to time", but this behavioral aspect is not described on other (especially higher) levels.”

This participant emphasizes the inconsistencies of the CEFR descriptors for phonological control, mentioning that some features appear in some descriptors but not in others and how this leads to a vague use of the scale. This response is aligned with other two answers discussing how the evaluators should be trained in order to use the scale. Rater 10 observed that

“In my opinion, the phonological control scale is a good source as it provides information necessary for teachers and test takers to understand what is expected from the students in each level. However, people who use the scale as reference should be well trained and informed about each of the descriptors, the technical terms and the theory underlying phonetics and phonology to be able to use the scale as it should.”

Rater 4, in their turn, acknowledged that “Yes, it could be. However, the person using the scale should be trained in using the scale and learn more about how such evaluation has to be carried out. Likewise, it would be essential to have great knowledge and experience with English phonology.” These three participants, raters 13, 10 and 4, bring out a reflection on how the descriptors are better used when used by professionals or trained raters. This study precisely selected teachers of ESL to be the evaluators because they supposedly have a background knowledge in assessment and would make better use of the scales in comparison to a naive listener. Nonetheless, results would be even more accurate if the raters went through a training process specific to the CEFR scale for phonological control. Due to time constraints this was not possible, but it is definitely valid for the raters to make this inference.

Other answers corroborate with our second hypothesis, which stated that the CEFR scale for phonological control is a helpful and applicable tool to the process of intelligibility and comprehensibility assessment, based on the raters’ report on their experience using the phonological control scale. Raters 9, 7, 2, 1, and 14, respectively, mentioned that “I believe it can be a good source to assess pronunciation, but I felt some details were missing when using it to evaluate the listenings.”; “There is no doubt that it is a great guide for teachers to assess students pronunciation, but there is room for improvement.”; “Yes, it is. It tries to guide teachers on how to evaluate students' production.”; “I believe it is a good resource to assess pronunciation, but not to assess proficiency.”; and “Yes, I do. It seemed to be a valid tool.”. Although some of them believe the descriptors could be better, they all agree that they are a useful tool for assessment.

The next question focused on these possible improvements for the phonological control scale descriptors, by asking the participants their suggestions. Rater 12 explained that

“Maybe it could be lack of my own attention, but as I was rating, I really missed something related to "mispronounced a phoneme and produced another word instead, that could change the meaning of the sentence, unless the interlocutor makes an effort to grasp from the context". I say that because I am a teacher and

researcher, and because of that, I could always understand what the participant might have intended to say when he mispronounced...but what about the people he will encounter in real life contexts? I believe it could cause serious communication problems. I also missed something about hesitations..I don't know if is more related to speaking overall, but I guess too many hesitations really affect comprehensibility!”

The report from this participant shows a valid suggestion; however, the descriptors already mention that, for beginner speakers, an effort may be required from the interlocutor in order to establish communication. The descriptor for A1 level, for instance, displays that the speaker “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)” (CEFR, 2018); for A2 level, as well, it is mentioned that

“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.” (CEFR, 2018)

It is assumed by the CEFR (2018) companion volume that any interlocutor is capable of understanding a beginner speaker of a second language, while making an effort or not, not only teachers, as suggested by rater 12.

Most of the answers for this question (do you think the phonological scale could be improved? how?) are concerning the lack of examples and details, as different participants pointed out. Rater 10, for instance, responded that “Yes, I believe the scale could provide speech samples so as to understand what is expected in each of the descriptors. These samples could be transcriptions, listening samples, examples of words/transcriptions, and so on.” Rater 9 highlighted the need for more details as well “Yes. Adding more details to each level. Some of the times I had to decide on what is closer to how I felt from the listenings, but it lacked on details to be truthful to the assessment.” Raters 4, 2 and 3, respectively, stated that “Yes. It could show some examples in order to illustrate the differences among the levels”; “Yes, I do. I believe there could be some examples and some definitions of specific terms.”; and “Yes, by having more specific information or examples that could guide the assessment”. It is understandable their need to have more details or even examples to guide

the assessment, as it could help them to provide a more accurate evaluation of the speakers' overall phonological control. Notwithstanding, it could be unmanageable to have such large descriptors with examples that could end up leading to a misguided judgment of the performance, because a single example cannot comprise all components permeating the phonological control of a speaker. All in all, these comments are highlighting the importance of training examiners to use the CEFR scale appropriately. During training sessions, raters can be provided with performance samples ranging from a variety of phonological control levels to help them understand the descriptors.

For the final question, "do you have any other observations?" some thought-provoking answers emerged. Rater 1, for instance, called attention to the sound quality of some samples:

"I'm afraid sound quality might have impacted my comprehensibility ratings and my transcription sometimes. P16, for example, had a very poor sound quality and I had a hard time understanding them. It was difficult for me to know if the source of the difficulty was in their pronunciation or in the sound quality, especially because it was hard to determine whether participants had accurately produced codas (plurals and third person conjugations, for instance) and because I had to pay extra attention to understand them."

Although the quality of the sounds were indeed improved through the software Audacity, as mentioned in the method section, there were some recordings that remained not clear enough. The data of this study was collected during the Covid-19 pandemic and thus the speech samples were remotely recorded, through Zoom, with the participants using their own microphones. The sound quality would definitely be better if data were collected in a lab with the appropriate equipment. Having this in mind, the ratings could have undoubtedly been affected and results could have been different if the sound quality was better.

Rater 7 acknowledged that having the same language background as the speakers (Portuguese as a first language) impacted the evaluation:

"Since their accent was so familiar I don't think I had problems in understanding them. Although I could tell they did vary in levels of proficiency, if we only take into consideration the extent to which their speech was intelligible (regardless of syntax, lexicon and morphology), basically all of them showed the same level of phonology control."

This is a very rich observation and it is in dialogue with Bent and Bradlow (2003) and Delatorre (2017) regarding the “matched interlanguage speech intelligibility benefit”. These studies indicated that a non-native speech is more intelligible to a non-native interlocutor sharing the same first language, as highlighted by rater 7 when explaining that “since their accent was so familiar I don’t think I had problems in understanding them”.

Raters 10 and 13 mentioned the complexity of separating phonological control from other features of speech when evaluating; they said, respectively, that “As a teacher, it was hard for me to evaluate and level speakers solely based on their pronunciation. It was a bit hard not to evaluate them as a whole.” and “It's very hard to dissociate phonological control from other proficiency variables (like grammar) when rating a speaker. I believe this takes some training.”

This section presented and discussed the responses from the questionnaire answered by the raters. There were open and closed questions aiming to gather as much information as possible from their experience using the CEFR phonological control scale as a tool to assess intelligibility and comprehensibility. All of the seven questions from the questionnaire were discussed in this section while trying to interpret the raters’ responses focusing on the second research question of this study: “To what extent is the CEFR scale for phonological control helpful in the assessment of intelligibility and comprehensibility by experienced L2 teacher raters?”

The responses that emerged from this questionnaire were intriguing and very helpful for our analysis, as we could make inferences with the literature review and the quantitative results. There were many responses requesting improvements to the scale, explaining difficulties using the descriptors and suggesting revisions; the literature acknowledges that these interferences, such as difficulty of the task, seriousness of the rater, criterias and the scale being used, have an impact on the final result (TOFFOLI; ANDRADE; BORNIA; QUEVEDO-CAMARGO, 2016). Nevertheless, there were as well many responses praising the phonological control scale and describing its usefulness as a tool for the assessment of intelligibility and comprehensibility. Thus, we can consider that the listeners tend to have a positive view of using the CEFR phonological scale, but perhaps the lack of training sessions to prepare them to use the scale generated difficulties to use the instrument and can account for some criticism and suggestions for improvement.

4.4 Summary

This chapter presented the results of this study while trying to answer the research questions. Quantitative and qualitative results were presented and investigated, making reference to other studies in the area. The results show that there is a highly significant correlation between intelligibility and comprehensibility, suggesting that highly intelligible speakers are easier to comprehend; moreover, there was a highly significant correlation between comprehensibility and phonological control, meaning that comprehensibility is a good predictor for raters when assigning a phonological control CEFR level. A moderate correlation was found between intelligibility and phonological control, which means that a speaker can be highly intelligible but still not be considered as a speaker possessing intermediate or advanced levels of phonological control, due to pronunciation patterns and possibly other confounding factors such as grammar and vocabulary knowledge. The responses from the qualitative results demonstrated usefulness for the phonological control scale descriptors, but not without, pointing out some inconsistencies in the descriptors and highlighting the need to provide users with specific training to learn how to use the instrument for assessment. The next chapter will conclude this thesis with final remarks, limitations of the study, pedagogical implications and suggestions for further research.

5. CONCLUSION

The objectives of this study were (a) to investigate how the intelligibility and the comprehensibility constructs are incorporated in the CEFR phonological scale (b) to compare the performance of experienced raters when using measures of intelligibility and comprehensibility and the CEFR scale, and (c) to correlate the CEFR phonological control scale with intelligibility and comprehensibility measures commonly used in L2 speech research. In order to achieve that, we collected data from 16 speakers from different university majors. A Speech Elicitation Task was administered to the speakers in which they had to orally describe an image. Their speech samples were recorded and used to create a listening task in a Google Form to go through the evaluation of the 14 raters participating in this study.

The listening tasks involved an orthographic transcription of the speech by the raters. These transcriptions were later checked by the researcher to check for accuracy and to generate an intelligibility score (percentage of words correctly transcribed). Furthermore, the listening task contained a Likert-scale for the listeners to assign a level of comprehensibility for each speaker, as well as the CEFR phonological control scale descriptors to place a level (A1-C2) for each speaker. Finally, the raters responded to a questionnaire with open and closed questions about their experience using the phonological control scale.

The research questions of this study were 1) how do raters' judgements of L2 learners' intelligibility and comprehensibility relate to the CEFR scale for phonological control? and 2) to what extent is the CEFR scale for phonological control helpful in the assessment of intelligibility and comprehensibility by experienced L2 teacher raters?. The first research question was answered through our quantitative analysis of the data generated by the listening task. The correlation coefficients displayed a highly significant relationship between intelligibility and comprehensibility, and between comprehensibility and phonological control as well; there was a moderate correlation between intelligibility and phonological control. Considering these significant correlations between the variables, we can affirm that our first hypothesis, there is a positive, significant correlation between intelligibility measures, comprehensibility ratings and the phonological control scores from the CEFR scale, suggesting that the new descriptors for phonological control adequately reflect the results of intelligibility and comprehensibility assessment for each proficiency level (A1-C2), was partially confirmed. We could not achieve a significant correlation between phonological control and the intelligibility variables, probably due to limited range of scores for the

intelligibility variable, given that all 16 speakers received high intelligibility scores (around 93% and 100%).

The second research question was investigated through the answers from the questionnaire. The raters of this study shared their opinion about the CEFR phonological control scale descriptors and some of them provided very detailed comments. We were able to partially confirm our second prediction that the CEFR scale for phonological control is a helpful and applicable tool to the process of intelligibility and comprehensibility assessment, based on the raters' report on their experience using the phonological control scale, as a result of some responses criticizing the descriptors and suggesting improvements, as well as highlighting the need to receive training on how to use the scale.

5.1 Limitations of the study and suggestion for further research

A remark often made by the raters in the questionnaire was concerning the difficulty in using the scale without receiving any training; even though the raters were all experienced teachers of English as an L2 with background in assessment, the fact of not receiving a 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 can be achieved by providing the participants with proper training sessions for them to get familiarized to the scale and to fully understand each descriptor, which is a suggestion for further research in the topic.

Another limitation was the amount of participants, speakers and listeners. Replicating the study with a bigger population may as well result in different correlation coefficients, for instance, and thus find different relationships between the 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 as a consequence of the Covid-19 pandemic and its restrictions.

5.2 Pedagogical implications

Intelligibility and comprehensibility were the main constructs observed in this research. The objective of the study was to understand the relationship between these two constructs with the assessment of the phonological control by raters using the CEFR (2018) scale. In what concerns 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 concerned with 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 results demonstrated that beginner speakers can be highly intelligible and able to have a conversation in their second language, despite being in the early stages of learning, and that should be the focus of teachers when dealing with the oral aspects of the language with their students. Nonetheless, these results have to be taken with some caution, given that in the present study, both speakers and listeners share the same L1, which means that communication may have been facilitated because the listeners (experienced English teachers) are familiar with the speakers' L2 pronunciation features.

REFERENCES

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

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

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>

DELATORRE, F. **Intelligibility of English Verbs Ending in -ed for Brazilian Learners of English Listeners**. Orientadora: Dra. Rosane Silveira. 2017. 288 f. Dissertação (Mestrado). Universidade Federal de Santa Catarina, Florianópolis, 2017.

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

DERWING, T., & MUNRO, M. Second Language Accent and Pronunciation Teaching: A Research-Based Approach. **TESOL Quarterly**, v. 39(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**, v. 29(3), p. 359-380. 2008.

DEYGERS, B., & VAN GORP, K. Determining the scoring validity of a co-constructed CEFR-based rating scale. **Language Testing**, v. 32(4), p. 521–541. 2015.

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**, v. 15(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 University Press (OUP). <http://dx.doi.org/10.1093/elt/ccs037>.

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**, v. 74(2), p. 253–278. 2018.

GONÇALVES, A. R. **In Search of Speech Intelligibility: The Case of English High Front Vowels**. Orientador: Rosane Silveira. 2014. 172 f. Dissertação (Mestrado em Inglês) - Universidade Federal de Santa Catarina, Florianópolis, 2014.

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

KANG, O., RUBIN, D., & KERMADE, A. Effect of training and rater individual differences on oral proficiency assessment. **Language Testing**, v. 36(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**, v. 68(1), p. 115–146. 2017.

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

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

MUNRO, M. J., & DERWING, T. M. Processing Time, Accent, and Comprehensibility in the Perception of Native and Foreign-Accented Speech. **Language and Speech**, V. 38(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**, v. 45(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**, v. 1(1), p. 11–42. 2015.

PALLANT, J. *SPSS Survival Manual: A step by step guide to data analysis using SPSS for Windows*. Allen & Unwin, 2005.

QUEVEDO-CAMARGO, Gladys. Breve história da evolução do construto proficiência em línguas. **Em Aberto**, Brasília, v. 32, n. 104, p. 27-44, jan. 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**, v. 7(2), p. 217–240. 2015.

SILVEIRA, R., & MARTINS, T. Assessing Second Language Oral Proficiency Development With Holistic And Analytic Scales. **Ilha Do Desterro**, v. 73(3), p. 227–250. 2020.

TOFFOLI, Sônia Ferreira Lopes; ANDRADE, Dalton Francisco de; BORNIA, Antonio Cezar; QUEVEDO-CAMARGO, Gladys. Avaliação com itens abertos: validade, confiabilidade, comparabilidade e justiça. **Educ. Pesqui**, São Paulo, v. 42, n. 2, p. 343-358, abr. 2016.

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

TROFIMOVICH, P., & ISAACS, T. Disentangling accent from comprehensibility. **Bilingualism**, v. 15(4), p. 905–916. 2012.

APPENDICES

Appendix A - Background Questionnaires (Speakers and Listeners)

Questionário de participação

Muito obrigada por participar desta pesquisa! Este formulário contém questões sobre seu uso e experiência com a língua inglesa. Após terminar o questionário, você deverá responder a este e-mail para agendar uma sessão na plataforma Zoom, onde você realizará um curto teste de proficiência em inglês, com o intuito de indicar seu nível. Por fim, você será convidado a descrever uma imagem em inglês, durante esta mesma sessão no Zoom. Você deverá levar em torno de 15 minutos para responder a este questionário.

Se tiver alguma dúvida, por favor entre em contato por e-mail ou telefone:

thaisy.sm@gmail.com

48991923095

***Obrigatório**

Nome: *

A sua resposta

E-mail: *

A sua resposta

Idade: *

A sua resposta

Cidade: *

A sua resposta

Curso de graduação: *

A sua resposta

Você possui algum nível de proficiência em outras línguas além do inglês? Se sim, quais? *

A sua resposta

Há quanto tempo você estuda inglês? *

A sua resposta

Onde você já estudou inglês? *

- Ensino primário
- Ensino fundamental
- Ensino médio
- Universidade
- Cursos de idiomas
- Curso no exterior

Quantas horas por dia você utiliza o inglês (falar, ouvir, ler e escrever)? *

- Menos de duas horas
- Entre duas e seis horas
- Entre seis e dez horas
- Mais de dez horas

Background Information

Age:

A sua resposta _____

City:

A sua resposta _____

How long have you been an English teacher?

A sua resposta _____

Where have you taught English? (middle school, high school, language school, university, etc)

A sua resposta _____

Appendix B - Oxford Placement Test

Oxford University Press
and
University of Cambridge Local Examinations Syndicate

Name:

Date:

quick placement test

Version 2

This test is divided into two parts:

Part One (Questions 1 – 40) – All students.

Part Two (Questions 41 – 60) – Do not start this part unless told to do so by your test supervisor.

Time: 30 minutes
Part 1

Questions 1 – 5

- Where can you see these notices?
- For questions **1** to **5**, mark **one** letter **A**, **B** or **C** on your Answer Sheet.

- 1

You can look, but don't touch the pictures.

A in an office
B in a cinema
C in a museum

- 2

Please give the right money to the driver.

A in a bank
B on a bus
C in a cinema

- 3

NO PARKING PLEASE

A in a street
B on a book
C on a table

- 4

CROSS BRIDGE FOR TRAINS TO EDINBURGH

A in a bank
B in a garage
C in a station

- 5

KEEP IN A COLD PLACE

A on clothes
B on furniture
C on food

Questions 11 – 20

- In this section you must choose the word which best fits each space in the texts.
- For questions 11 to 20, mark **one** letter **A, B, C** or **D** on your Answer Sheet.

Good smiles ahead for young teeth

Older Britons are the worst in Europe when it comes to keeping their teeth. But British youngsters (11) more to smile about because (12) teeth are among the best. Almost 80% of Britons over 65 have lost all or some (13) their teeth according to a World Health Organisation survey. Eating too (14) sugar is part of the problem. Among (15), 12-year olds have on average only three missing, decayed or filled teeth.

- | | | | | |
|----|-----------|----------|------------|-----------|
| 11 | A getting | B got | C have | D having |
| 12 | A their | B his | C them | D theirs |
| 13 | A from | B of | C among | D between |
| 14 | A much | B lot | C many | D deal |
| 15 | A person | B people | C children | D family |

Christopher Columbus and the New World

On August 3, 1492, Christopher Columbus set sail from Spain to find a new route to India, China and Japan. At this time most people thought you would fall off the edge of the world if you sailed too far. Yet sailors such as Columbus had seen how a ship appeared to get lower and lower on the horizon as it sailed away. For Columbus this **(16)** that the world was round. He **(17)** to his men about the distance travelled each day. He did not want them to think that he did not **(18)** exactly where they were going. **(19)**, on October 12, 1492, Columbus and his men landed on a small island he named San Salvador. Columbus believed he was in Asia, **(20)** he was actually in the Caribbean.

- 16** A made B pointed C was D proved
- 17** A lied B told C cheated D asked
- 18** A find B know C think D expect
- 19** A Next B Secondly C Finally D Once
- 20** A as B but C because D if

Questions 21 – 40

- In this section you must choose the word or phrase which best completes each sentence.
- For questions **21** to **40**, mark **one** letter **A**, **B**, **C** or **D** on your Answer Sheet.

- 21 The children won't go to sleep we leave a light on outside their bedroom.
A except B otherwise C unless D but
- 22 I'll give you my spare keys in case you home before me.
A would get B got C will get D get
- 23 My holiday in Paris gave me a great to improve my French accent.
A occasion B chance C hope D possibility
- 24 The singer ended the concert her most popular song.
A by B with C in D as
- 25 Because it had not rained for several months, there was a of water.
A shortage B drop C scarce D waste
- 26 I've always you as my best friend.
A regarded B thought C meant D supposed
- 27 She came to live here a month ago.
A quite B beyond C already D almost
- 28 Don't make such a! The dentist is only going to look at your teeth.
A fuss B trouble C worry D reaction
- 29 He spent a long time looking for a tie which with his new shirt.
A fixed B made C went D wore
- 30 Fortunately, from a bump on the head, she suffered no serious injuries from her fall.
A other B except C besides D apart

- 31 She had changed so much that anyone recognised her.
A almost B hardly C not D nearly
- 32 teaching English, she also writes children's books.
A Moreover B As well as C In addition D Apart
- 33 It was clear that the young couple were of taking charge of the restaurant.
A responsible B reliable C capable D able
- 34 The book of ten chapters, each one covering a different topic.
A comprises B includes C consists D contains
- 35 Mary was disappointed with her new shirt as the colour very quickly.
A bleached B died C vanished D faded
- 36 National leaders from all over the world are expected to attend the meeting.
A peak B summit C top D apex
- 37 Jane remained calm when she won the lottery and about her business as if nothing had happened.
A came B brought C went D moved
- 38 I suggest we outside the stadium tomorrow at 8.30.
A meeting B meet C met D will meet
- 39 My remarks were as a joke, but she was offended by them.
A pretended B thought C meant D supposed
- 40 You ought to take up swimming for the of your health.
A concern B relief C sake D cause

Part 2

Do not start this part unless told to do so by your test supervisor.

Questions 41 – 50

- In this section you must choose the word or phrase which best fits each space in the texts.
- For questions 41 to 50, mark **one** letter **A, B, C** or **D** on your Answer Sheet.

CLOCKS

The clock was the first complex mechanical machinery to enter the home, (41) it was too expensive for the (42) person until the 19th century, when (43) production techniques lowered the price. Watches were also developed, but they (44) luxury items until 1868 when the first cheap pocket watch was designed in Switzerland. Watches later became (45) available and Switzerland became the world's leading watch manufacturing centre for the next 100 years.

- | | | | |
|--------------|------------|-------------|------------|
| 41 A despite | B although | C otherwise | D average |
| 42 A average | B medium | C general | D common |
| 43 A vast | B large | C wide | D mass |
| 44 A lasted | B endured | C kept | D remained |
| 45 A mostly | B chiefly | C greatly | D widely |

Dublin City Walks

What better way of getting to know a new city than by walking around it?

Whether you choose the Medieval Walk, which will (46) you to the Dublin of

1000 years ago, find out about the more (47) history of the city on the Eighteenth

Century Walk, or meet the ghosts of Dublin's many writers on the Literary Walk, we know you will

enjoy the experience.

Dublin City Walks (48) twice daily. Meet your guide at 10.30 a.m. or 2.30 p.m. at

the Tourist Information Office. No advance (49) is necessary. Special

(50) are available for families, children and parties of more than ten people.

46 A introduce B present C move D show

47 A near B late C recent D close

48 A take place B occur C work D function

49 A paying B reserving C warning D booking

50 A funds B costs C fees D rates

Questions 51 – 60

- In this section you must choose the word or phrase which best completes each sentence.
- For questions 51 to 60, mark **one** letter **A, B, C** or **D** on your Answer Sheet.

- 51 If you're not too tired we could have a of tennis after lunch.
 A match B play C game D party
- 52 Don't you get tired watching TV every night?
 A with B by C of D at
- 53 Go on, finish the dessert. It needs up because it won't stay fresh until tomorrow.
 A eat B eating C to eat D eaten
- 54 We're not used to invited to very formal occasions.
 A be B have C being D having
- 55 I'd rather we meet this evening, because I'm very tired.
 A wouldn't B shouldn't C hadn't D didn't
- 56 She obviously didn't want to discuss the matter so I didn't the point.
 A maintain B chase C follow D pursue
- 57 Anyone after the start of the play is not allowed in until the interval.
 A arrives B has arrived C arriving D arrived
- 58 This new magazine is with interesting stories and useful information.
 A full B packed C thick D compiled
- 59 The restaurant was far too noisy to be to relaxed conversation.
 A conducive B suitable C practical D fruitful
- 60 In this branch of medicine, it is vital to open to new ideas.
 A stand B continue C hold D remain

Appendix C - Listening Tasks

Proficiency Assessment

Speakers were recorded describing an image as much detailed as they could. There are speakers with different levels of proficiency. We selected the first 40 seconds of their descriptions to be assessed by you in terms of intelligibility and comprehensibility.

There are 4 main steps:

- 1) Transcribe every word you hear from the speech samples (do not "correct" their speech when transcribing; it is important to transcribe the sample with as much precision as you can);
- 2) Select on the Likert scale the level of difficulty you had when comprehending their speech;
- 3) Assign each speaker a CEFR level from the phonological control scale (A1-C2);
- 4) Answer a questionnaire about your experience using the phonological control scale.

You will be presented with the image that is being transcribed and with the phonological control scale for you to consult. There are 10 participants. We recommend the use of headphones when listening to the speech samples. You can listen to the samples as many times as you find necessary.

Thank you for participating in this research! If you have any questions, please contact thaisv.sm@gmail.com

Participant 1

(audio sample)

(image being described)

Transcribe every word you hear from the speech sample. *

A sua resposta

How difficult was it to understand the speech? *

1 2 3 4 5 6 7 8 9

no difficulties

extremely difficult

Read the phonological control scale and assign one level for the participant *

PHONOLOGICAL CONTROL			
	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 four points of teacher message are clear and precise. Intelligibility and effective conveyance of 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 of 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 of the sounds of the target language, some features of accent retained from other language(s) may be noticeable, but they do not affect intelligibility.	Can articulate virtually all of the sounds of the target language with a high degree of control. He/she can usually self-correct if he/she reliably mispronounces 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 he/she means to say.
B2	Can generally use appropriate intonation, place stress correctly and articulate individual sounds clearly. Accent tends to be influenced by other language(s) he/she speaks, 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 he/she is aware to predict the phonological features of most utterance words (e.g. word stress) with reasonable accuracy (e.g. whilst reading).	Can employ prosodic features (e.g. stress, intonation, rhythm) to support the message he/she intends to convey, though with some influence from other languages he/she speaks.
B1	Pronunciation is generally intelligible; can approximate intonation and stress at both sentence and word levels. However, accent is usually influenced by other language(s) he/she speaks.	Is generally intelligible throughout, despite regular mispronunciation of individual sounds and words he/she is less familiar with.	Can convey teacher message in an intelligible way in spite of a strong influence on stress, intonation and/or rhythm from other language(s) he/she speaks.
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 other language(s) he/she speaks on stress, rhythm and intonation may affect intelligibility, requiring clarification from interlocutor. 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 other language(s) he/she speaks. Prosodic features (e.g. word stress) are adequate to 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 concerned. Can reproduce correctly a limited range of sounds as well as the stress on 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 other language(s) he/she speaks. Teacher interlocutor needs to be collaborative.

Page 136 • CEFR Companion Volume with New Descriptors

You can access the image for better reading here <https://bit.ly/3mYQGo7>

- A1
- A2
- B1
- B2
- C1
- C2

Report your experience using the CEFR's Phonological Control Scale

How difficult it was to use the scales?

1 2 3 4 5 6 7 8 9

no difficulties extremely difficult

Do you agree with the descriptors for each level?

- Yes
- No

Do you think the descriptors are too detailed?

- Yes
- No

Do you think the descriptors lack in detail?

Yes

No

Do you think the phonological control scale is a good source to assess pronunciation? Why?

A sua resposta _____

Do you think the phonological control scale could be improved? How?

A sua resposta _____

Do you have any other observations?

A sua resposta _____

Appendix D - Transcriptions

S1

in the picture i see what seems to be an office, in the ceiling there are four light bulbs, in the wall i see a clock, and four frames, i see two bookshelves with books and in the right bookshelf there is also a little plant, on the floor on the left i see a water filter with a little cup waiting to be filled, on the right there is a plant, in the middle of the picture i see four adults in the table

S2

in the image there is a meeting between four people, the girl sitting in the left of the table is wearing a yellow sweater, the guy standing in the left is wearing a purple sweater, the girl right - in the right side of the table, standing, is wearing a pink shirt and a purple skirt

S3

it is a room, it looks like an office, it has a lot of different kinds of blue in the colors, there are some people, two woman and two man, and they are having coffee and one of them has a computer. there are some books, there are four light bulbs, there are one clock, there are a few books, and a plant

S4

i'm seeing four people, they are colorful, the way that they are dressed, there is a woman, two women and one man having a cup of coffee, i guess, they seem to be in a library

S5

a lot of people, probably a family, a lot of blue, the girl who uses a yellow shirt also uses a - i don't know - yellow shoe, they are a happy family and they are drinking some coffee probably

S6

there are people drinking coffee, sitting on a table, they seem to be working in a coffee shop or an office, there is a water place to drink water, and books on the wall, and some plants and they seem happy

S7

i see in this picture four people meeting and there are two men and two women, and one women is sit in a chair, with, a glass, a cup of coffee and a notebook, there is a man sit too, and he's drinking coffee, smiling. and a couple is stand, they are talking, there is a blue background, there is water

S8

so, this seems to be a conference room, maybe a break - actually a break room where people are taking a break from their jobs, so there are four people in this image, two women and two men, they are drinking coffee and talking, there are one table and two chairs, the women are drinking coffee in a mug, there is a man drinking coffee as well, and they seem to be laughing and enjoying their breaks

S9

so, there are four people, two men, two women, one woman is drink something, maybe coffee, tea, i dont know, and the same time she is using the computer, and we have a man and he is also sit, i dont remember sentado if i can say sit

S10

to describe this picture, i see four people, three of them are drinking coffee, one of them bought coffee from the coffee shop, i see a table, two chairs, a couple of books, a water machine, the time in the clock on the wall is five-fifty, i think, there are a couple plants, like a cactus, and apparently, this is an office, because

S11

so there is an image of a cafe of some sorts with four people in it, two of them are sitting in chairs, drinking coffee, one of them with a mug, the other one with a cup of coffee, there's two people they're standing up talking with each other, one of them are - is pointing with their hands towards the other person, and the other one has a glass of coffee and is also pointing to the other person

S12

i see four people, two stand up, i dont know if this word is right, a lot of blue, shades of blue, perspective in the window, i think it's a window, one clock, books, water

S13

there are four people, two woman and two men, the woman in the right, or in the left, the one who is sitting, he wears a yellow shirt, and the other one wears a pink shirt and a skirt, the group is talking and smiling and drinking coffee, they are probably at work

S14

i think this image shows a - some colleagues at their office, during their, not lunchtime but like, i forgot the word, but their free time during their work day, they are talking about what, i dont know, stuff from their lives, some of them were working there before, this girl with the yellow shirt, she was working before

S15

it is a workspace, with a window, a clock, couple of bucks, four person, two womans, two mans, they are getting coffee, taking coffee, drinking coffee, they are drinking coffee, and the one of the woman has a book, notebook on the table

S16

i can see a group of people, i believe they're working on something, maybe actual work, or it's something related to school, but there's two people sitting down and two people standing up, they are working on something in the computer, they are wearing really colorful clothes, there's two woman, women and two men

Appendix E - CEFR Phonological Control Scale

PHONOLOGICAL CONTROL			
	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 his/her message are clear and precise. Intelligibility and effective conveyance of 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 retained from other language(s) may be noticeable, but they do not affect intelligibility.	Can articulate virtually all of the sounds of the target language with a high degree of control. He/she can usually self-correct if he/she noticeably mispronounces 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 he/she means to say.
B2	Can generally use appropriate intonation, place stress correctly and articulate individual sounds clearly; accent tends to be influenced by other language(s) he/she speaks, 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 his/her repertoire to predict the phonological features of most unfamiliar words (e.g. word stress) with reasonable accuracy (e.g. whilst reading).	Can employ prosodic features (e.g. stress, intonation, rhythm) to support the message he/she intends to convey, though with some influence from other languages he/she speaks.
B1	Pronunciation is generally intelligible; can approximate intonation and stress at both utterance and word levels. However, accent is usually influenced by other language(s) he/she speaks.	Is generally intelligible throughout, despite regular mispronunciation of individual sounds and words he/she is less familiar with.	Can convey his/her message in an intelligible way in spite of a strong influence on stress, intonation and/or rhythm from other language(s) he/she speaks.
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 other language(s) he/she speaks 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 other language(s) he/she speaks. 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 concerned. Can reproduce correctly a limited range of sounds as well as the stress on 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 other language(s) he/she speaks; his/her interlocutor needs to be collaborative.

Appendix F - Original scores provided by the 14 raters

Intelligibility

S	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14
S1	100	100	100	98	98	100	98	100	97	100	96	97	98	96
S2	96	94	100	100	100	98	98	100	100	96	94	94	100	100
S3	100	96	96	100	98	98	96	100	98	98	98	100	100	100
S4	94	97	94	97	91	86	88	97	100	97	97	94	94	100
S5	100	94	100	97	94	94	94	100	89	94	100	94	100	94
S6	100	100	100	100	100	100	100	100	97	100	97	100	100	100
S7	100	96	98	98	100	96	93	100	96	98	98	96	98	100
S8	98	100	100	98	100	100	98	100	100	100	97	100	100	100
S9	100	93	100	95	97	97	95	97	100	97	100	100	100	100
S10	98	100	96	98	96	100	98	96	98	93	93	96	100	96
S11	98	93	100	97	94	98	96	96	100	98	98	98	98	98
S12	97	100	100	97	97	100	100	97	97	100	94	97	100	97
S13	100	100	100	100	100	100	100	100	100	100	98	100	100	100
S14	100	100	100	94	98	98	98	100	98	98	98	100	100	100
S15	97	100	100	97	100	100	100	100	97	97	97	97	97	97
S16	98	100	98	98	98	96	96	100	100	98	94	100	100	100

Comprehensibility

S	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14
S1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
S2	1	3	1	2	5	2	2	3	1	1	3	2	1	1
S3	1	1	3	2	1	3	3	1	1	1	2	1	2	2
S4	5	2	5	5	8	5	3	5	2	2	5	3	5	2
S5	1	5	1	4	6	5	3	3	1	2	3	3	2	2
S6	2	1	1	2	1	1	3	1	3	2	4	1	1	1
S7	3	3	2	2	6	2	4	3	3	2	5	1	2	1
S8	1	1	1	1	1	1	3	1	1	1	3	1	1	1
S9	2	1	1	4	2	1	3	2	1	2	4	2	2	1
S10	3	5	2	3	2	4	2	2	5	3	5	3	2	1
S11	1	6	1	4	2	2	3	4	2	1	3	2	2	2
S12	1	6	1	3	1	1	3	1	1	1	5	2	2	1
S13	2	1	1	3	1	2	2	1	1	2	3	1	2	1
S14	1	4	1	3	2	2	3	1	5	4	4	1	3	1
S15	1	5	3	3	1	2	3	3	2	5	6	2	3	3
S16	4	1	1	3	2	1	3	1	2	2	5	1	3	1

Phonological Control

S	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	
S1	5	5	6	5	5	4	5	5	5	5	5	5	5	4	5
S2	4	3	5	4	3	2	4	3	2	4	4	3	3	3	3
S3	5	6	3	4	6	3	4	5	5	5	4	5	4	4	3
S4	4	4	2	3	1	2	4	2	2	3	3	3	3	3	3
S5	4	3	6	3	2	1	4	1	3	2	4	3	3	3	3
S6	4	4	5	4	5	4	4	2	2	2	3	5	4	4	4
S7	4	3	4	3	3	3	3	2	2	1	3	5	3	3	3
S8	4	5	1	4	5	4	4	5	5	5	3	5	5	5	5
S9	3	3	4	2	4	3	4	2	3	1	3	3	2	2	2
S10	4	3	4	2	5	3	4	4	2	4	3	2	4	3	3
S11	4	3	5	3	5	4	4	3	3	4	4	4	4	4	3
S12	4	3	5	2	5	2	4	2	2	1	3	2	2	2	1
S13	4	5	4	3	5	2	5	3	3	1	4	5	3	2	2
S14	4	4	5	3	5	3	4	5	2	1	3	5	3	3	3
S15	4	2	3	2	5	2	4	1	2	1	2	2	2	2	2
S16	4	5	5	3	5	4	4	5	2	3	3	5	3	3	3