DOI: http://dx.doi.org/10.5007/2175-8026.2020v73n3p227

ASSESSING SECOND LANGUAGE ORAL PROFICIENCY DEVELOPMENT WITH HOLISTIC AND ANALYTIC SCALES

Rosane Silveira[1*]

Thaisy da Silva Martins[1**]
[1]Universidade Federal de Santa Catarina, Santa Catarina, Florianópolis, Brasil

Abstract

The present study discusses how experienced raters use different types of scales to assess the development of oral proficiency in English as a second language (L2). Raters assigned rates to speech samples first using a holistic scale (CEFR, 2018) and then assigning rates for pronunciation, vocabulary, grammar and fluency performance using individual scales. The speech samples were recorded by five Brazilians. There were two data collection sessions, with the second one occurring 7-8 months after the first one. The results indicate high levels of agreement among raters for all scales. Furthermore, the raters detected changes in speakers' performance in four out of five scales: L2 oral proficiency, vocabulary, grammar, and fluency, and these differences in rates across time were significant for oral proficiency, vocabulary, and fluency. Thus, the different types of scale allow detecting L2 oral proficiency development.

Keywords: L2 Assessment; Oral Proficiency Development; Holistic and Analytic Scales.

[*] Associate Professor at the Department of Foreign Language and Literature and at the English Graduate Program (UFSC). CNPq Research Fellow. Coordinator of the English Graduate Program (2019-current). Coordinator of the NUPFFALE Research Group (Núcleo de Fonética e Fonologia Aplicadas à Língua Estrangeira). E-mail: rosanesilveira@hotmail.com. ORCID: https://orcid.org/0000-0003-0329-0376.

[**] Master's student in the English Graduate Program (PPGI) at the Federal University of Santa Catarina (UFSC), member of the NUPFFALE Research Group (*Núcleo de Pesquisa em Fonética e Fonologia Aplicadas à Língua Estrangeira*). E-mail: thaisy.sm@gmail.com. ORCID: https://orcid.org/0000-0002-5638-9102.

1. Introduction

Whether one is engaged in second language research or second language teaching, a pressing concern is to understand the language proficiency construct. Two relevant questions to pursue this goal are: What are the components of L2 proficiency? What is the best way to assess L2 proficiency? In his review of research articles in the field of bilingual language cognition, Hulstijn (2012) highlights the fact that language proficiency is a variable that has been investigated either as an independent or a moderator variable, and that a number of instruments are used to assess this construct: vocabulary tests, such as the Peabody Picture Vocabulary Test or Nation's vocabulary tests; translation tests; mean length of run measures[1]; panels of judges; fluency measures; self-rating questionnaires; or even years or levels of language-program enrolment.

One important point raised by Hulstijn (2012) is that language proficiency can be assessed in terms of basic language cognition and/or high language cognition. The author explains that basic language cognition involves linguistic knowledge (representation and use of information) that all native speakers of a given language with no language impairment are assumed to possess, and it is restricted to oral language (listening, speaking) in utterances with high-frequency elements. On the other hand, high language cognition encompasses extended linguistic knowledge that is gained with the development of literacy skills, age, formal education and professional/personal experiences, and it includes the processing of written language (reading and writing) in sentences that may contain low-frequency elements.

In the present study, we attempt to discuss how second language (L2) proficiency can be measured, for research purposes, with a focus on what Hulstijn (2012, 2015) calls basic language cognition, which he also refers to as the core components of L2 proficiency. The core components consist of linguistic cognition "in the phonetic-phonological, morpho-phonological, morpho-syntactic, and lexical/pragmatic domains" (Hulstijn, 2015, p. 42).

As Shin (2013, p. 1) observes, it is essential for second language researchers to find ways to define and describe "what it means to know a language and to use it for communicative purpose". In other words, the field needs to have a proper definition for proficiency, as well as adequate measures that match the definition.

In this study, we make reference to the proficiency levels proposed by the *Common European Framework of Reference for Languages* (CEFR). Launched in 2001, the guidelines provided by the CEFR aim at providing a common basis for language syllabus design, curriculum guidelines, examinations and textbook design. In its 2018 version, the CEFR guidelines define proficiency as "a term encompassing the ability to perform communicative language activities (can do…), whilst drawing upon both general and communicative language competences (linguistic, sociolinguistic, and pragmatic), and activating appropriate communicative strategies" (CEFR - Companion Volume with New Descriptors, 2018, p. 32). As Shin (2013) explains, the CEFR adopts a "real-life"

approach to proficiency, meaning that proficiency is defined as "what language users can do with language in communicative tasks in real life" (pp. 1-2).

The present study aims at investigating whether L2 oral proficiency development can be assessed by experienced raters using a holistic scale and four analytic scales. The holistic scale proposed by the CEFR is originally composed of seven descriptors in the form of "Can do" statements, six of which are associated with the proficiency levels proposed by the CEFR. These levels are Proficient User: C2 (Mastery), C1 (Effective Operational Proficiency); Independent User: B2 (Vantage), B1 (Threshold); Basic User: A2 (Waystage), A1 (Breakthrough). The seventh descriptor indicates lack of proficiency. The analytic scales were designed to assess important subcomponents of proficiency, namely pronunciation, vocabulary, grammar, and fluency, all of which using a 10-point scale. In this study, we compare raters' performance when they evaluate speech samples produced by Brazilian learners of English using these different types of scale to examine whether raters detect changes in L2 oral proficiency development in a formal instruction context.

2. Assessing L2 Oral Proficiency

When assessing L2 oral proficiency for research purposes, authors have used different types of elicitation tasks and have investigated different types of measures. In this section, we will illustrate these different methodological options by reviewing relevant studies that are concerned with the investigation of L2 speech development and that rely on longitudinal data.

Aiming to investigate developmental aspects of adult immigrant ESL learners' speech production, Derwing, Munro and Thomson (2007) focus on fluency and comprehensibility development, which are seen as important aspects for communicative success. In their study, comprehensibility is defined as "ease or difficulty with which a listener understands L2 accented speech" (Derwing et al., 2007, p. 360), and it is measured with a Likert-scale, which reflects the perception of ordinary listeners, not second language specialist. The authors believe that the comprehensibility measure is useful for tracing L2 learners' progress and assessing pedagogical intervention. As for fluency, it is defined as "an automatic procedural skill on the part of the speaker and a perceptual phenomenon in the listener (Derwing et al., 2007, p. 360), and it is measured by intuitively analyzing features of dysfluency (filled pauses, excessive pausing, pausing in inappropriate places, false starts, and slow speaking rate), using a Likert-scale.

Derwing et al. (2007) intended to examine how comprehensibility and fluency develop over a two-year period, and how these traits are related to learners' exposure to English outside the classroom. They examined data provided by two groups of immigrants: sixteen native speakers of Mandarin and sixteen native speakers of Slavic languages. The participants were all adult learners of English, enrolled in full-time beginner English classes in Canada. The participants were recorded in seven data collection sessions, in which they always produced a narrative based on a set of pictures. Additional data related to English learning

experiences and uses were gathered with the use of a questionnaire and a semi-structured interview.

Three 20-second speech excerpts[2] from the opening of the narratives (produced by each participant) were presented to a group of native speakers of English, volunteers enrolled in an undergraduate Education program. This group of listeners had some familiarity with the first language (L1s) spoken by the L2 learners. The listeners rated the L2 learners' comprehensibility and fluency using a holistic seven-point Likert-scale, with 1 assigned to speech that was considered extremely easy to understand or extremely fluent, and 7 assigned to speech that was deemed extremely difficult to understand or extremely dysfluent.

Comprehensibility results showed that both the Mandarin and the Slavic speakers received intermediate rates in the first speech sample rated by the English raters (overall rates for Time 2 = 4.2 in the seven-point scale). The second speech sample was collected after ten months the study had started (Time 6) and the overall rate was 4.0. There was a different trend for the two groups of speakers: the Mandarin speakers received slightly worse rates than they had received for the first speech sample, while the Slavic speakers received slightly better rates. In the third speech sample, collected at the end of the study (Time 7), the overall rate was 3.6, thus indicating the speech produced was deemed to be easier to understand. Both groups of speakers received similar rates again, thus showing that the Mandarin speakers' performance was considered easier to comprehend this time, while the Slavic speakers remained with rates similar to the ones they received for the second speech sample. A similar trend was observed for the fluency results, but the overall rates were harsher for fluency than for comprehensibility (Time 2 = 4.8, Time 6 = 4.5, and Time 7 = 4.1).

The statistical analysis showed that the Mandarin group had no significant improvement over time, while the Slavic group improved somewhat (significant differences were found for the comprehensibility and fluency measures when comparing Time 2 to Time 6, and Time 2 to Time 7). The correlational analysis using the comprehensibility and fluency rates yielded strong, positive correlations. The authors attempt to explain the difference in performance for the two groups by discussing how these groups' experiences with L2 use could be related to the fact that the Slavic group obtained a better performance due to their frequent interactions with English speakers and willingness to communicate with people from another culture.

With the goal of investigating how different interactional conditions (monologic vs. dialogic tasks) impact L2 speech production over time, Ferrari (2012) employs three measures that are traditionally used in the field to evaluate production data: syntactic complexity, syntactic accuracy, and speech fluency. The author designed a longitudinal study to investigate the development of these traits over time and across four tasks for L2 learners and native speakers of Italian. The study took place in a three-year period and had four moments of data collection. Data were gathered from four L2 learners from different L1 backgrounds who had been living in Italy for a minimum period of four years and from two Italians, all between 15 to 19 years of age. When the study started, the L2 learners' proficiency

level was B1 or B2, according to the CEFR proficiency levels. The tasks used to collect speech samples were two monologic tasks (film picture and story retelling) and two interactive tasks (semi-structured interview and telephone call).

From the three measures used by Ferrari (2012), the present study is concerned with fluency, which is the only measure that is intrinsically related to speech production, given that syntactic complexity and syntactic accuracy are measures that can be used for the assessment of both written and speech production. In Ferrari's (2012) study, fluency is defined as real-time language processing and it is measured in terms of number of silent pauses and number of hesitation phenomena (filled pauses and functionless repetitions) per speech units.

Ferrari (2012) concludes that fluency improves in the long run for L2 learners, especially when data from the first and the fourth data collection sessions are compared. Furthermore, monologic tasks lead to less fluent speech, while dialogic tasks generate highly fluent speech, and these task-type effects are observed for both L2 learners and native speakers of Italian, although they are more prominent in the four sessions of the native speakers' productions, while for L2 learners, task effects are more prominent when data from the first and the fourth data collection are compared.

Saito and Hanzawa (2016) conducted a study to examine how and to what extent English formal instruction enables adolescent and adult learners to improve their L2 oral abilities. More specifically, the study focused on accentedness and phonological phenomena such as segmental accuracy, word stress and intonation accuracy, and fluency, which was measured in terms of speech rate. Accenteness is defined as "how different an L2 speaker's accent sounds from that of the native-speaker community (Saito & Hanzawa, 2016, p. 820). The study included a set of independent variables commonly investigated in the field of second language learning: length and focus of instruction, frequency of L2 conversation, aptitude, and motivation.

Saito and Hanzawa (2016) analyzed extemporaneous speech samples produced by 56 Japanese learners of English, ages ranging from 18 to 19 years, all pursuing a major in a Japanese college. All the Japanese participants reported having never studied English abroad and having had a few hours of English classes per week in the regular school for about six years. The study also included baseline data provided by experienced Japanese speakers of English residing in Canada for over 20 years and who used English for communication on a daily basis. All speakers completed a timed-image-description task, which consisted of descriptions of seven unrelated images with three keywords added at the bottom of each image. The descriptions produced for the last three images were evaluated. The inclusion of keywords was meant to minimize disfluencies, while the timed task was meant to control the amount of speech monitoring.

Accentedness was measured by asking listeners' with no language teaching experience to assign impressionistic ratings. Five English speakers rated 66 speech samples using a sliding scale ranging from 0 to 1000 (very negative and very positive evaluation of accentedness, respectively). The speech samples consisted of 5-10 seconds excerpts extracted from each of the three images used to collect the speech samples (average of 25s of speech for each speaker).

For the evaluation of segment, prosody and fluency measures, five experienced raters (graduate students from an English program in a Canadian university) were recruited and trained. These raters were native speakers of English and were trained to perform subjective judgments of the pronunciation and fluency aspects of L2 speech. The experienced raters listened to 66 samples and rated each of them using the sliding scale and 1 was assigned to productions that were nontarget-like, while 1000 was assigned to target-like productions. Four variables were generated by the experienced raters: vowel and consonant errors, word stress errors, intonation inappropriateness, speech rate inappropriateness.

The variables used as predictors for oral L2 performance were gathered with a questionnaire, which was used to obtain information about the Japanese speakers' length of English instruction inside and outside the foreign language (FL) classroom, focus of FL instruction, amount of second language use, and motivation. Furthermore, the Japanese speakers completed the LLAMA language aptitude test (Meara, 2005). The data generated by these instruments were correlated with the oral production data.

The results for the oral proficiency variables showed that the Japanese students' group performed significantly differently from the group of experienced Japanese speakers of English (baseline data). The overall accentedness rates were 287 for the Japanese students and 806 for the experienced Japanese speakers. Higher rates were assigned for the phonological measures. The Japanese students got the following rates: segmental errors = 411, word stress errors = 496, intonation = 439, and speech rate = 496. The experienced Japanese speakers obtained rates that ranged from 841 (segments) to 927 (speech rate). Thus, it is clear that the raters distinguished between the performance of the two groups, identifying the experienced Japanese speakers as more proficient than the Japanese students, with a few Japanese students receiving rates in the range obtained by the experienced Japanese speakers.

As for the variables obtained with the help of the questionnaire and the LLAMA aptitude test, the authors combined them to create six factors that were used to predict the participants' performance on the L2 oral measures. From these six factors (Factor 1: recent and extra FL experience; Factor 2: regular FL classroom experience; Factor 3: academic motivation; Factor 5: amount of L2 use; Factor 6: professional and integrative motivation), only Factor 1 came out significant. Factor 1 was a composite variable including length of FL instruction outside the classroom, pronunciation training and amount of time communicating with non-native speakers. Thus, for this study, additional exposure to the L2, which came from formal instruction outside the regular classroom and extensive communication with non-native speakers of English, predicts performance on accentedness, segmentals, and word stress measures.

De Jong, Groenhout, Schoonen (2015), and Hulstijn (2015) investigated whether fluency measures such as pause duration, number of filled pauses, and syllable duration, which are deemed to be influenced by personality traits and speaking style, are valid predictors of L2 proficiency. The authors used scores from

an L2 vocabulary test as an estimate of L2 proficiency. Then, they checked whether a number of fluency measures, with and without correction for speakers' style using L1 fluency measures, could predict performance on the L2 proficiency measure. The authors question the validity of the fluency measures used in L2 research, as they seem to ignore the fact that people have different speaking styles in their L1, which need to be taken into consideration when measuring L2 fluency.

In their study, De Jong et al. (2015, p. 224) define L2 fluency as a subcomponent of L2 proficiency, regarding it as "speedy and smooth delivery of speech without (filled) pauses, repetitions, and repairs". In De Jong et al. (2015, p. 226), fluency was measured as number of silent pauses, non-lexical filled pauses, repetitions, and corrections, and length of silent pauses, and mean duration of syllables.

The authors collected data from English and Turkish learners of Dutch, including fluency measures from their L1 and L2, as well as an L2 proficiency measure (vocabulary test). They designed eight speaking tasks for L2 and a corresponding version for the participants' L1s, varying in degree of formality, complexity, and discourse mode (persuasion and description). The results indicated that L1 fluency measures can be a good predictor for L2 fluency measures. As for correcting L2 fluency measures by using L1 fluency outcomes, De Jong et al. (2015) found that a corrected measure of syllable duration is a better predictor of L2 proficiency than the uncorrected measure, although both measures succeed in predicting L2 proficiency.

As can be seen from the previous discussion, assessing oral proficiency in a second language is a complex task. So far we have discussed some of the different variables assessed in L2 research, yet another source of debate is the role played by human raters who are often in charge of assessing L2 oral proficiency (e.g. Bent & Bradlow, 2003; Cruz, 2008; Derwing & Munro, 1997; Gass &Varonis, 1984; Hayes-Harb & Hacking, 2015; O'Brien, 2016; Shintani, Saito & Koizumi, 2017; Silveira & Silva, 2018). Research has addressed two main issues concerning raters' differences: L1 background and experience with L2 accented speech.

For the present study, it is important to highlight the findings of Huang and Jun (2015), who report that stricter rates were assigned by inexperienced native speakers, while experienced native speakers and experienced non-native speakers rated the participants in a very similar manner and were more often able to identify native speakers. Thus, Huang and Jun (2015) conclude that experienced non-native speakers perform similarly to experienced native speakers when completing a task to rate foreign language accent. These findings help to validate the participation of advanced learners of English who are also English teachers as raters in L2 research.

As this brief review of empirical research concerned with the development of L2 speech has shown, authors have used a number of elicitation tasks and oral proficiency measures. In this study, we opted for an elicitation task that is deemed adequate for low-proficiency speakers (image-description task). This task was used to gather speech samples to obtain an oral proficiency measure based on the CEFR guidelines. Additionally, we were interested in investigating how experienced raters perform when using different types of scales to assess

speech data, so we decided to add four scales that assess subcomponents of L2 proficiency, namely, pronunciation, vocabulary, grammar, and fluency.

3. Objectives and Research Questions

The present study investigates the assessment of L2 oral proficiency development by a group of raters using holistic descriptors from the CEFR, as well as analytic scales to rate the speakers' performance in terms of pronunciation, vocabulary, grammar, and fluency accuracy. The objective is to investigate whether experienced raters can detect L2 oral proficiency development when using different types of scales.

The study aims to answer the following research question: How do experienced raters assess L2 oral proficiency development in a formal instruction context using different types of scales?

4. Method

In this section, we will provide relevant information about the participants who acted either as speakers or listeners (raters) in the present study. We will also describe the research instruments used to collect data from each group of informants, and the procedures for data collection and analysis. The data presented here is part of a larger research project that has been submitted to the Ethics in Research Board at *Universidade Federal de Santa Catarina* (CAAE: 55740116.6.0000.0121,) and it has been approved (Review number: 1.597.582).

4.1. Speakers

The study includes speech samples from five Brazilian learners of English. These samples were collected in two different sessions, separated by a period of seven or eight months. Table 1 summarizes relevant biographical information about the speakers. All the speakers were undergraduate students, one attending the English *Letras* program, and four attending the *Secretariado Executivo* program at a university located in the south of Brazil. These were first-year students who attended an average of 240 hours of English classes per year, and these classes were part of their undergraduate program curriculum. Four participants were female and one was male, and their ages ranged from 18 to 32 (mean = 22.8). Three of these learners were from the city of Florianópolis, in the state of Santa Catarina, one of them was from Santarém, Pará, and one from Ampére, Paraná.

Table 1: Speakers' background

Participant	Major	Age	Sex	Place of birth
P1	Secretariado	18	F	Florianópolis – SC
P2	Letras Inglês	18	F	Santarém – PA
P3	Secretariado	21	m	Ampére – PR

P4	Secretariado	25	f	Florianópolis – SC
P5	Secretariado	32	f	Florianópolis – SC

Source: the authors.

In addition to the data provided by the first-year students, the study includes speech samples produced by a male native speaker of American English and two advanced learners of English. These data are merely used to check for rater reliability (see section 5).

4.2. Raters

As Table 2 shows, three raters contributed to the study by evaluating the speech produced by the Brazilian learners. The three of them are experienced English teachers with at least ten years of experience (mean = 16.6 years). All raters have high education levels, with two pursuing a doctoral degree in an English graduate program, and one who completed a Bachelor's of Arts degree. Two of the raters are Brazilians and one is Australian. Two of the raters reported having intermediate or advanced knowledge of other L2s. All the raters were residing in Florianópolis by the time of data collection.

Table 2: Raters' Background

	Sex	Age	Homestate	Education	Other L2	Teaching
Rater 1	M	25	Paraná	Doctoral student	Spanish, French	10 years
Rater 2	F	34	Toowomba (Australia)	BA	Portuguese	10 years
Rater 3	F	48	Santa Catarina	Doctoral student		30 years

Source: the authors.

4.3. Instruments and procedures to collect speech samples

The speakers were invited to participate in the study and recorded other pronunciation tasks that will not be analyzed in this article. Here the focus is to discuss oral proficiency development and how this construct is assessed by raters.

Data from the speakers were collected using the following instruments: Questionnaires, an Image Description Test (English proficiency), and a Proficiency Assessment Form. Prior to the start of data collection, a consent form was presented to the speakers, so that they became aware of the general objectives of the research and how the data would be collected. Participants interested in contributing to the research signed the consent form. Subsequently, they answered a background questionnaire. For the speakers, the questionnaire sought to obtain important information, such as age, gender, city of origin, Brazilian Portuguese variety, educational level, foreign language knowledge, learning experiences, and use of the English language. An Image Description Test, adapted from Silveira (2011, 2012)

was used for the assessment of the speakers' English oral proficiency level. The test contained five slides with images that were not necessarily related. The informants were instructed to look at the pictures and describe them in English as they were recorded. An initial 30-second sample of each informant's speech was selected to be assessed by the experienced English teachers who acted as raters.

Each speaker attended two individual data collection sessions, with the second session taking place seven or eight months after the first one. The sessions started with the presentation of the consent form, followed by the completion of the questionnaire. Then, an assistant researcher tested the recording equipment[3] and explained the procedures for recording the Image Description Test. The test contained five slides with images that were not necessarily related. Informants were instructed to look at the pictures and describe them in English as they were recorded. The participants took from 3 to 10 minutes to complete the task. An initial 30-second sample of each informant's speech was selected to be rated. The audio files were edited using the Audacity software to remove low-frequency noise and to normalize the samples to the same peak level.

The raters also answered a questionnaire used to gather biographical data and information about their experiences as English teachers. The raters completed a Proficiency Assessment Form, which included the following information: audio files for each speaker; an adapted version of the Overall Speech Production Scale provided by the CEFR (2001, p. 58); four rating scales to assess the speakers' pronunciation, vocabulary, grammar and fluency (scale range: 1 (poor performance) to 10 (excellent performance), and an optional open-ended question to comment on each speaker's performance[4]. As it can be seen in Chart 1, the Overall Production Scale included seven descriptors, six of which correspond to the CEFR proficiency levels, which range from C2 (the descriptor at the top) to A1. The descriptor at the bottom indicates lack of proficiency. In the form provided to the raters, the information about the proficiency levels was omitted.

Chart 1: Adapted version of the CEFR Overall Production Scale.

..... Can produce clear, smoothly flowing well-structured speech with an effective logical structure which helps the recipient to notice and remember significant points.
..... Can give clear, detailed descriptions, developing particular points.
..... Can give clear, systematically developed descriptions, with appropriate highlighting of significant points, and relevant supporting detail.
..... Can give clear, detailed descriptions on a wide range of subjects, expanding and supporting ideas with subsidiary points and relevant examples.
..... Can reasonably fluently sustain a straightforward description of one of a variety of subjects presenting it as a linear sequence of points.
..... Can give a simple description as a short series of simple phrases and sentences linked into a list.
..... Can produce simple mainly isolated phrases about people and places.

Source: Adapted from the CEFR Guidelines (2001, p. 58).

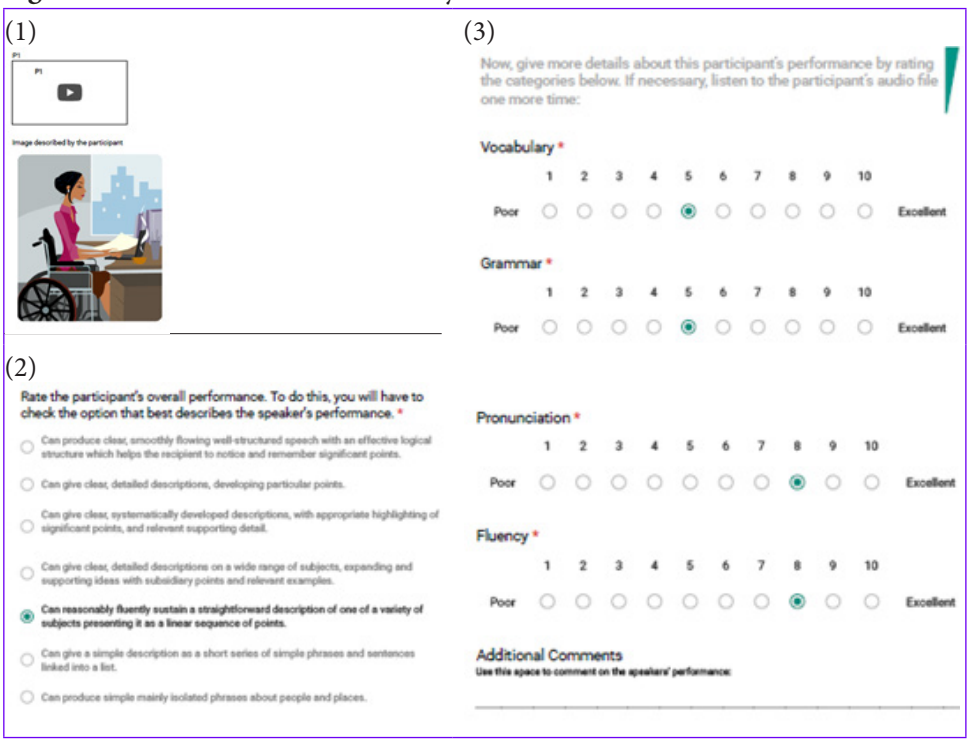
The audio files were saved as video files so that they could be used to prepare the Pronunciation Assessment Form using Google Forms. The form was designed to allow raters to perform the evaluations individually and at their chosen pace. First, the raters completed the biographic questionnaire, which was also made available through Google Forms. Then, they were asked to wear headphones and initiated a practice session to familiarize themselves with the descriptors, the analytic scales, and the purpose of the proficiency assessment task. The raters received written instructions on how to perform the task, which were organized into three steps:

1. STEP 1: Listen to each participant speak for about 30 seconds and rate the participants' overall performance. To listen to the participants, open the audio file for each participant.

- STEP 2: Rate the participant's overall performance. To do this, you will have to check (✔) the option that best describes the speaker's performance.
- STEP 3: Provide specific information about vocabulary, grammar, pronunciation and fluency.
- Now, give more details about this participant's performance by rating the categories below. If necessary, listen to the participant's audio file one more time.

For each speaker, the rater would see a form like the one in Figure 1, displayed on the computer screen, containing the speech sample followed by the image being described (1), the holistic scale with descriptors (2), and the analytic rating scales (3).

Figure 1: Screenshot of the Proficiency Assessment Form.



Source: the authors.

In addition to the data being analyzed here, which was produced by first-year undergraduate students, the form contained data from one English native speaker, two advanced learners of English, and other English learners who completed a single recording session and, for this reason, had their data discarded from this study. The native speakers' and the advanced learners' data were used to monitor raters' behavior. Each rater evaluated a total of 23 speech samples: two speech samples produced by five English learners (10), plus one speech sample produced by one native speaker (1), one speech sample produced by two advanced learners of English (2), and one speech sample produced by ten speakers who recorded a single session (10). Raters took about forty minutes to complete the task and, once they were finished assigning all ratings, they sent the form through Google Forms. All forms were downloaded and the ratings were organized in Excel and SPSS for analysis.

5. Data Analysis

First, the overall L2 oral proficiency ratings assigned to each of the five speakers were organized in a spreadsheet, identifying the ratings that were assigned to the speech samples from the first and the second data collection sessions. For this data, each of the seven CEFR descriptors made available to the raters were converted into numbers, which resulted in a scale ranging from 1 (low proficiency) to 7 (high proficiency, or C2 level).

To check for rater reliability, the ratings were submitted to interrater analysis using Cronbach Alpha before calculating the mean rate for each speaker. The ratings assigned to the native speaker and the advanced English learners were inspected to check for rater reliability as well.

The second step was to organize the ratings assigned with the pronunciation, vocabulary, grammar and fluency scales in different spreadsheets. These scales already contained numbers ranging from 1 to 10, and they were simply organized to identify the two different data collection sessions (Time 1 and Time 2). Again, the ratings assigned to the native speaker and the advanced learners were inspected to check for rater reliability. The data generated after steps one and two were organized in tables and analyzed as follows: (a) calculation of Cronbach Alpha for each dependent variable to determine interrater reliability, (b) inspection of results to identify how the speakers were ranked by the raters, (c) calculation of descriptive statistics (mean, standard deviation [SD], minimum and maximum rates), and (d) comparison of ranks for Time 1 and Time 2 (given the small sample size, Wilcoxon Tests were used to compare group ranks).

The third step consisted of running correlations among all the dependent variables using the ratings assigned to the five speakers. Two-by-two correlations were run using the variables overall English oral proficiency, pronunciation, vocabulary, grammar and fluency ratings. Given the small sample size, Spearman correlations were run. All statistical analyses were performed with SPSS (20.0), with alpha level established at .05.

An important issue regarding rater reliability is whether they are capable of distinguishing different proficiency levels using the scales provided in this study. For this purpose, the data collection instruments included speech samples produced by one English native speaker and two advanced learners of English. All raters succeeded in assigning high rates to these three speakers. The average rate for overall English oral proficiency reached 6.55 (maximum rate possible = 7), while the average rates for pronunciation, fluency, grammar, and vocabulary ranged from 9.75 to 9.91 (maximum rate possible = 10). These results confirm that the three raters identified the three speakers as highly proficient, with the native speaker receiving slightly higher rates than the advanced learners. Note that the raters were not warned that the data set contained speech samples from one native speaker or from two advanced learners. Having described the research method, the next section will present and discuss the research results.

6. Results and Discussion

English speech samples produced by first-year undergraduate students enrolled in the *Secretariado Executivo* Program and the English *Letras* Program were rated by three experienced raters. The curriculums of both programs include extended hours or English classes. To trace L2 oral proficiency development, the speech samples were recorded when the students were attending the first semester of the program and when they were concluding the second semester (7 or 8 months after the first data collection).

In order to answer the research question, we will discuss the rates assigned to each speaker, beginning with the results obtained with the CEFR overall L2 oral proficiency scale. We begin by presenting the results of the interrater reliability analysis, which is followed by a discussion of how the raters ranked the participants according to their performance in the two data collection moments (Time 1 and Time 2). The next step is to analyze the rates obtained with the vocabulary, pronunciation, grammar and fluency scales, following the same procedures adopted for the overall L2 oral proficiency variable.

The rates assigned to English oral proficiency were submitted to interrater analysis to determine how much raters agreed on their ranking of the five speakers. The Cronbach Alpha analysis showed a high reliability rate for the two data collection times (Time 1: Cronbach $\alpha = .92$; Time 2: Cronbach $\alpha = .89$). In other words, the raters reached over 79% of agreement as how each speaker was ranked across time.

Having established high interrater reliability, the mean rates for overall English oral proficiency were calculated. The five speakers received higher rates in Time 2, especially P1 and P3. Raters agreed that P2 and P1 are the most proficient speakers and that P4 is the least proficient.

Table 3: Overall Proficiency rates for the two data collection sessions

Time 1				
	Rater 1	Rater 2	Rater 3	Mean rate
P1	3.0	3.0	3.0	3.0
P2	6.0	4.0	7.0	5.6
P3	3.0	2.0	2.0	2.3
P4	1.0	1.0	1.0	1.0
P5	2.0	3.0	2.0	2.3
				Mean: 2.86 SD: 1.72
Time 2				
P1	6.0	3.0	3.0	4.0
P2	6.0	5.0	7.0	6.0
P3	3.0	3.0	4.0	3.3
P4	2.0	1.0	1.0	1.3
P5	3.0	3.0	2.0	2.6
				Mean: 3.46 SD: 1.72

Rating scale ranged from 1 (limited proficiency) to 7 (high proficiency).
Source: the authors.

The results displayed in Table 3 show lower mean proficiency ratings for Time 1 (mean = 2.86, SD = 1.72) than Time 2 (mean = 3.46, SD = 1.72), indicating that the raters detected a better speaker performance across time. In order to check for differences in ranks of overall English oral proficiency across Time 1 and Time 2, a Wilcoxon test was run. The results show that the differences observed were significant ($Z = -2.07$, $p = .03$), thus suggesting that after 7-8 months of formal instruction, a significant improvement in overall oral proficiency occurred.

Having presented the results for the overall L2 oral proficiency measure, we turn now to the results for the analytic scales that were used to assess the development of pronunciation, vocabulary, grammar, and fluency across time.

The interrater reliability analysis for the pronunciation measure showed that the raters achieved high reliability levels (Cronbach $\alpha = .90$ for Time 1 and .96 for Time 2). In other words, there was over 81% of agreement among raters. Table 4 shows the individual pronunciation rates assigned by each rater, as well as the mean rate for each speaker across time. Only P1 and P3 showed improvement across time, P2 received the maximum score across time, while P4 and P5 received lower pronunciation ratings for Time 2 than for Time 1. P2 obtained the highest ranks, and P4 consistently occupied the lowest ranks across time. It is interesting to note that the mean pronunciation rates actually remain the same across time (Time 1: mean = 6.80, SD = 2.08; Time 2: mean = 6.80, SD = 2.87), thus indicating that pronunciation performance did not change across time. The Wilcoxon test used to compare the ranks across time showed no significant

differences (Z = .00, p = 1), which confirms that the speakers' pronunciation did not change significantly after 7-8 months attending English classes.

Table 4: Pronunciation ratings

Time 1				
	Rater 1	Rater 2	Rater 3	Mean rate
P1	7	9	6	7.3
P2	10	10	10	10
P3	7	7	6	6.6
P4	4	7	3	4.6
P5	5	8	3	5.3
				Mean: 6.80 SD:2.08
Time 2				
P1	8	8	8	8
P2	10	10	10	10
P3	8	9	8	8.3
P4	3	4	2	3
P5	4	7	3	4.6
				Mean: 6.80 SD: 2.87

Rating scale ranged from 1 (poor performance) to 10 (excellent performance).
Source: the authors.

Turning to the vocabulary measure, the interrater reliability analysis showed strong agreement among the three raters for Time 1 (Cronbach α = .94) and for Time 2 (Cronbach α = .93), with a minimum of 83% agreement. As Table 5 shows, in Time 1, P2 and P1 received the highest vocabulary rates, while P4 and P5 received the lowest rates. In Time 2, P2 and P3 received the highest rates, while P4 and P5 received the lowest ones. When we compare the results for Time 1 and Time 2, we can see that P2, P3, P4, and P5 received higher scores in Time 2, with P3 showing greater improvement, while P1 remains with the same rate across time.

Table 5: Vocabulary ratings

Time 1				
	Rater 1	Rater 2	Rater 3	Mean rate
P1	7	7	6	6.6
P2	9	8	10	9
P3	7	5	5	5.6
P4	4	3	1	2.6
P5	4	5	3	4
				Mean: 5.60 SD: 2.44

Time 2				
P1	7	6	7	6.6
P2	10	9	10	9.6
P3	8	8	8	8
P4	3	4	2	3
P5	4	7	3	4.6
				Mean: 6.40 SD: 2.63

Rating scale ranged from 1 (poor performance) to 10 (excellent performance).
Source: the authors.

Table 5 displays the mean vocabulary rates for Time 1 (mean = 5.60, SD = 2.44) and Time 2 (mean = 6.40, SD = 2.63), thus indicating an increase in the rates for Time 2. Again, a Wilcoxon test was run to compare the ranks across time, yielding a non-significant difference (Z = -1.84, p = .06). Note, however, that the Wilcoxon test approached significance, which indicates that the speakers improved their performance overall, especially P3, but this improvement was not enough to yield significant results across time.

For the grammar measure, slightly more disagreement among the raters for Time 1 data was observed (Cronbach α = .78, 60% agreement). For Time 2, the interrater reliability coefficient was again very strong (Cronbach α = .96, 92% agreement), showing a high level of agreement. For Time 1, Rater 2 was displaying a different behavior when rating the speakers, which resulted in weak two-by-two correlation coefficients when paired with the other raters. Table 6 shows that P2 and P1 obtained the highest grammar rates for Time 1, and that P4 received the lowest rates. In Time 2, P2 and P3 received the highest rates, and P4 continued to be the speaker with the lowest grammar rates. We can see that three speakers received higher rates in Time 2 (P1, P2, and P3), while the other speakers received lower rates when compared to Time 1. One result that calls attention is the high rate that Rater 2 assigned to P5 in Time 1 (9). The discrepancy is pretty high, given that the other raters assigned 2 and 4 to P5 in Time 1.

Table 6: Grammar ratings

Time 1				
	Rater 1	Rater 2	Rater 3	Mean rate
P1	6	6	5	5.6
P2	9	8	10	9
P3	6	5	5	5.3
P4	4	3	1	2.6
P5	4	9	2	5
				Mean: 5.53 SD: 2.26

Time 2				
P1	7	6	7	6.6
P2	10	9	10	9.6
P3	9	8	8	8.3
P4	2	4	1	2.3
P5	5	5	3	4.3
				Mean: 6.26 SD: 2.96

Rating scale ranged from 1 (poor performance) to 10 (excellent performance).
Source: the authors.

In Table 6, we can observe that the mean grammar rates slightly increased across time (Time 1: mean = 5.53, SD = 2.26; Time 2: mean = 6.26, SD = 2.96), but the Wilcoxon test came out not significant (Z=-1.08, p=.27). This lack of significant improvement across time is probably due to the fact that two participants got lower rates in Time 2 (P4 and P5), two got slightly better rates, and only one improved a lot (P3).

Turning now to the fluency measure, the interrater reliability analysis showed strong agreement between the raters for Time 1 (Cronbach α = .94) and Time 2 (Cronbach α = .92), with a minimum of 84% of agreement. As it can be seen in Table 7, P2 and P1 received the highest rates for Time 1 and Time 2, while P4 received the lowest ones across time. For the first time, all speakers improved their rates across time, especially P3.

Table 7: Fluency ratings.

Time 1				
	Rater 1	Rater 2	Rater 3	Mean rate
P1	7	8	6	7
P2	10	8	10	9.3
P3	6	6	4	5.3
P4	3	3	1	2.3
P5	4	6	3	4.3
				Mean: 5.66 SD: 2.65
Time 2				
P1	8	8	8	8
P2	10	9	10	9.6
P3	8	7	8	7.6
P4	5	3	2	3.3
P5	6	7	3	5.3
				Mean: 6.80 SD: 2.47

Rating scale ranged from 1 (poor performance) to 10 (excellent performance).
Source: the authors.

When we look at the mean fluency rates across time (Table 7), we observe improvement (Time 1: mean= 5.66, SD = 2.65; Time 2: mean = T2: 6.80 , SD = 2.47), and a significant difference in ranking was found using the Wilcoxon test (Z = -2.06, p=.03). This confirms that the five speakers improved fluency performance after a period of 7-8 months attending English classes.

To conclude the results section, we will present the multiple correlation analysis including the five dependent variables investigated in this study: overall English oral proficiency, pronunciation, vocabulary, grammar, and fluency, splitting the data for each data collection time. This analysis is expected to show how the holistic scale used to rate L2 proficiency and the four analytic scales used to rate proficiency subcomponents are related.

Table 8 shows the Spearman correlation coefficients and probability (p) values for the four analytic measures when they are correlated with overall L2 oral proficiency, using data from Time 1 and from Time 2 separately. As can be seen, from the four correlations run using Time 1 data, one (overall L2 oral proficiency and vocabulary) came out as a perfect positive correlation (Spearman correlation = 1, p = .000). In other words, all the raters assigned the same ranks to the same speakers in these two scales, which means that overall L2 oral proficiency could be 100% of the time predicted by the variable vocabulary in Time 1. As for the other analytic scales, all of them yielded a strong positive and significant correlation with overall L2 proficiency in time 1 (.97, p = .005), meaning that pronunciation, grammar, and fluency could also predict the L2 proficiency performance 94% of the time. Similar outcome for the correlation analysis using Time 2 data was obtained, as can be seen in Table 8. Again, strong, positive, and significant correlations were obtained, but this time, all the analytic scale measures yielded strong, positive, and significant correlations with overall L2 oral proficiency (.90, p = .037), meaning that four analytic measures could predict the overall L2 oral proficiency ratings 81% of the time.

Table 8: Correlations between the five measures across time.

	Time 1			
	Vocabulary	Grammar	Pronunciation	Fluency
L2 Oral Proficiency	1,000	.975	.975	.975
	.000	.005	.005	.005
	Time 2			
	,900	,900	,900	,900
	,037	,037	,037	,037

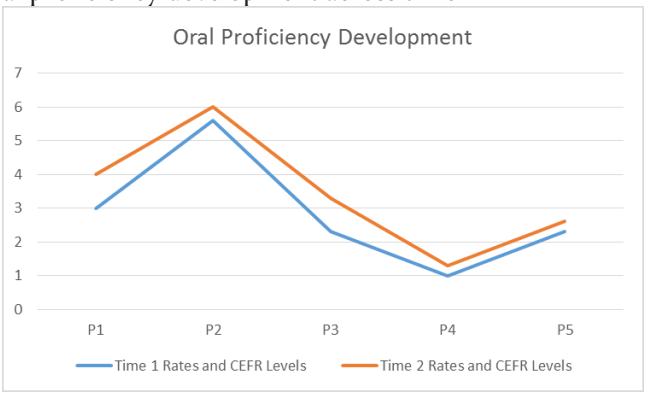
Source: the authors.

The results show that raters detected significant proficiency development within a 7-8 month period in a formal classroom environment for all five speakers. Previous studies with immigrants and learners enrolled in study-

abroad programs have shown a "window of opportunity" for development for the first year. In the present study, we were investigating L2 development in a formal setting, where opportunities to use the L2 are less frequent than when learners are in a country where the language is spoken on a daily basis. Even so, proficiency significant development was captured using both a holistic scale to assess English oral proficiency and one out of four analytic scales that addressed L2 proficiency subcomponents, namely, fluency performance.

First-year undergraduate students attending extended hours of English classes were rated as having varied L2 oral proficiency levels both at the onset of data collection (Time 1) and at the final data collection (Time 2). As Figure 2 shows, the proficiency levels at Time 1 ranged from 1 for P4 to 5.6 for P2. In Time 2, rates ranged from 1.3 (P4) to 6 (P2). In other words, the speakers' oral proficiency level ranged from limited proficiency (P1) to intermediate/advanced (P2). This figure also shows that there was improvement for all speakers in Time 2, but that P1 and P3 are the ones whose rates increased the most.

Figure 2: Oral proficiency development across time



Source: the authors.

Figure 3 illustrates L2 development for the four analytic measures, which are subcomponents of L2 proficiency. A clear developmental path for the five speakers is observed for vocabulary and fluency, with most participants displaying improvement in performance across time, which is similar to the results for the overall L2 oral proficiency measure (Figure 2). Again, P3 is the speaker who displays larger increases in the vocabulary and fluency rates across time. However, when we look at the graphs for pronunciation and grammar, we see a different scenario, which is mainly due to the performance of P4, who got lower rates at Time 2 than at Time 1. For these two measures, we see some improvement for P1, P2 (grammar, because P2 got maximum rate for pronunciation at both Time 1 and Time 2), P3, and P5.

It is interesting to notice that significant changes in rates across time were found for L2 oral proficiency, fluency only, with vocabulary approaching significance. These results suggest that changes in pronunciation and grammar seem to be more resilient and may take more time and different pedagogical

strategies to be consolidated, and that at least one of the speakers (P4) seems to be in need of more systematic help to improve her pronunciation and grammar skills.

Figure 3: Development of Pronunciation, Vocabulary, Grammar and Fluency across time.

Source: the authors.

Answering the research question, we can state that both the holistic and the analytic scales employed in this study are consistently used by the raters to assess L2 proficiency development, and can, therefore, be considered good tools to investigate L2 proficiency. In this study, we followed the tradition of Derwing and Munro (2015) research by adopting Likert-scales to assess subcomponents of oral proficiency. Listeners' rates have been criticized for relying on intuition, thus being highly subjective. However, we agree with Derwing and Munro (2015) that listeners' rates using Likert-scales are a good option for assessment if we are concerned with the intelligibility of the speech produced by the language learners, rather than with the accuracy level (native-like performance) of L2 speech.

All analytic scales were highly correlated with the holistic scale used by the raters in this study. All in all, the raters displayed strong interrater agreement rates, with an exception being the rates assigned to the grammar measure at Time 1, which caused some disagreement among raters.

By examining the raters' behavior, we can corroborate Huang and Jun (2015) observation that experienced raters, no matter if they are native speakers or advanced learners of the L2 being analyzed, are equally competent in using different types of scale to assess proficiency and its multiple subcomponents.

7. Final Considerations

While investigating the appropriateness of holistic and analytic scales to assess L2 oral proficiency and some of its subcomponents, our study indicated strong positive correlations among the holistic measure and the four analytic measures. We also saw that each of the three raters were highly consistent in their use of the five measures, and in the ranks they assigned to the five speakers.

As for oral speech development, we saw that the five scales used in this study detected some degree of development for the speakers, but that each of these speakers presented different learning paths both at the initial and final data collection. These findings highlight the importance of examining L2 speech as it evolves, paying attention to the individual paths, rather than simply grouping all learners and drawing generalized conclusions about how L2 proficiency develops (Ortega & Han, 2017).

Although overall English oral proficiency development was found to be significant across time, the analytic scales showed significant change for fluency only. Vocabulary development approached significance and seems to be noticeable after 7-8 months of formal English instruction. However, pronunciation and grammar changes are hardly seen for the group of speakers being tested, with a few exceptions. These results may be partially due to the limited data provided to the raters and the small sample size. However, they still indicate that in formal language settings such as the one investigated here, where the communicative approach for language teaching prevails, special attention may need to be given to the teaching of grammar and pronunciation so that the development of these subcomponents can be enhanced.

Acknowledgements:
We thank CNPq for the financial support to conduct this study. We also thank the undergraduate students who were part of the study, Larissa Lannes for the help with the proficiency forms, and Laura Viana for the help with data collection.

Notes

2. Measure used to assess linguistic productivity, traditionally calculated by collecting 100 utterances and dividing the number of morphemes by the number of utterances.
3. Speech samples from the following data collection moments: Time 2 (two months), Time 6 (ten months), and Time 7 (two years).
4. All data collection sessions for the speakers took place at FONAPLI – Laboratório de Fonética Aplicada. This lab is equipped with a sound isolation booth, a *C 520 L* professional head-worn condenser microphone connected to the audio interface *MOTU Ultra Lite mk3* and an audio editor software, *Ocean Audio,* mono connected. All this equipment was connected to an iMac computer used to record the data.
5. Only Rater 2 added some comments about how a few participants resorted to Portuguese to complete the image description task and about the limited proficiency of a few speakers.

References

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

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*.http://www.coe.int/t/dg4/linguistic/sourceframework_en.pdf.

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment: Companion volume with new descriptors*. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Cruz, N. C. (2008). Familiaridade do ouvinte e inteligibilidade da pronúncia de aprendizes brasileiros de inglês. *Revista Horizontes de Linguística Aplicada*, 7(1), 88-103.

De Jong, N., Groenhout, R, & Schoonen, R. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, 36(2), 223-243.

Derwing, T. & Munro, M. (1997). Accent, comprehensibility and intelligibility: Evidence from four L1s. *Studies in Second Language Acquisition,* 19, 1-16.

Derwing, T. & Munro, M. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. (Vol. 42). Netherlands: John Benjamins Publishing Company.

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2007). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, 29(3), 359–380.

Ferrari, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development. In Housen, A., Kuiken, F., & Vedder, I. *Dimension of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 277-298). Amsterdam/Philadelphia: John Benjamins.

Gass, S. & Varonis, E. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, 34(1), 65–89.

Hayes-Harb, R. & Hacking, J. (2015). Beyond rating data: What do listeners believe underlies their accentedness judgments? *John Benjamins Publishing Company*, 1(1), 43-64.

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

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, 15(2), 422 – 433.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. Amsterdam: John Benjamins.

Meara, P. (2005). *LLAMA Language Aptitude Tests*: The Manual. Swansea: Lognostics. http://www.lognostics.co.uk/tools/llama/llama_manual.pdf.

Ortega, L. & Han, Z. (Eds.). (2017). *Complexity theory and language development: In Celebration of Diane Larsen-Freeman*. Amsterdam: John Benjamins.

O'Brien, M. (2016). Methodological choices in rating speech samples. *Studies in Second Language Acquisition,* 38(3), 587-605.

Saito, K. & Hanzawa, K. (2016). Developing second language oral ability in foreign language classrooms: The role of the length and focus of instruction and individual differences. *Applied Psycholinguistics*, 37, 813–840.

Shin, S.Y. (2013). Proficiency scales. In Chappelle, C.A. (Ed.). *The encyclopedia of applied linguistics*. (pp. 1-7). Oxford, UK: Wiley-Blackwell.

Shintani, N., Saito, K., & Koizumi, R. (2017). The relationship between multilingual raters' language background and their perceptions of accentedness and comprehensibility of second language speech. *International Journal of Bilingual Education and Bilingualism*, 22(7), p. 849-869.

Silveira, R. (2011). Pronunciation instruction and syllabic-pattern discrimination. DELTA - *Documentação de Estudos em Linguística Teórica e Aplicada*, 27(1), 5-22.

Silveira, R. (2012). L2 production of English word-final consonants: The role of orthography and learner profile variables. *Trabalhos em Linguística Aplicada*, 51(1), 15-28.

Silveira, R.; Silva, T. C. (2018). L2 speech intelligibility: Effects of coda modification, degree of semantic information and listeners' background. *Revista Brasileira de Linguística Aplicada*, v. 18, n. 3, p. 639-664.

Recebido em: 08/03/2020
Aceito em: 20/05/2020