Raters' Judgments Reliability in the Spotlight: Assessment Challenges in Pronunciation Research within ESP Teaching

Leticia Quesada Vázquez (lequesad@ucm.es)
0000-0002-7422-6251
Universidad Complutense de Madrid

Abstract: Many researchers nowadays study pronunciation as a means to improving second language students' communicative skills. Native listeners' perception tasks are considered a reliable tool although their veracity has been questioned due to the listeners' subjectivity and/or the type of task performed. The present study investigates the effectiveness of explicit rhythm instruction to Spanish/Catalan bilinguals who were engineering undergraduates to improve their comprehensibility and fluency in English. For this purpose, seven native speakers of English rated the extemporaneous speeches of forty-two students before and after instruction via a 5-point scale. Students were divided into an experimental group, receiving explicit rhythm instruction, and a control group, which did not. Unexpectedly, results showed that the experimental group was less comprehensible after treatment, and that the control group improved more than the experimental group in terms of fluency. However, previous results obtained by different acoustic analyses suggested further improvement in experimental students' learning process. Hence, the use of a mixed-method assessment is encouraged to support human judgments.

Keywords: English for Specific Purposes (ESP), English as a Foreign Language (EFL), pronunciation teaching, rhythm instruction, pronunciation assessment, raters' judgments

La fiabilidad de los juicios de los evaluadores en el punto de mira: Desafíos de la evaluación en la investigación sobre la pronunciación en la enseñanza del inglés para fines específicos

Resumen: Actualmente, muchos investigadores estudian la pronunciación como medio para mejorar las habilidades comunicativas de estudiantes de segundas lenguas. Las pruebas de percepción con oyentes nativos se consideran una herramienta fiable, aunque su veracidad ha sido cuestionada debido a la subjetividad de los oyentes y/o al tipo de tarea realizada. El presente estudio investiga la eficacia de la instrucción explícita de ritmo en estudiantes de ingeniería bilingües español/catalán para mejorar su comprensión y fluidez en inglés. Para ello, siete hablantes nativos de inglés calificaron los discursos improvisados de cuarenta y dos estudiantes antes y después de la instrucción mediante una escala de 5 puntos. Los estudiantes se dividieron en un grupo experimental, que recibió instrucción explícita de ritmo, y un grupo de control, que no la recibió. Sorprendentemente, los resultados mostraron que el grupo experimental fue menos comprensible después del tratamiento, y que el grupo de control mejoró más que el grupo experimental en términos de fluidez. Sin embargo, resultados previos obtenidos mediante diferentes análisis acústicos sugirieron una mayor mejora en el proceso de aprendizaje de los estudiantes del grupo experimental. Por lo tanto, se recomienda el uso de una evaluación de métodos mixtos para respaldar los juicios humanos.

Palabras clave: Inglés para fines específicos, inglés como lengua extranjera, enseñanza de la pronunciación, instrucción de ritmo, evaluación de la pronunciación, juicios humanos

1. Introduction

How to assess intelligibility, comprehensibility, fluency or accentedness progress in pronunciation teaching research has recently caused a heated debate. After the emergence of Levis' (370) *Intelligibility Principle* as the main focus for second language instruction, pronunciation teaching researchers are currently relying on different tools that go beyond pure acoustic analysis to assess how effectively English as a Second Language (ESL) learners are making themselves understood. In fact, Derwing and Munro (7) renounce acoustic analysis, since intelligibility is essentially conceived by both the listener's perception and the speaker's production. Therefore, the communicative dimension of this type of assessment leads us to the use of evaluative tools that imply human judgments.

The most popular technique applied is the listeners' appraisal by means of tests that range from forced choice identification tasks and transcription exercises to examine intelligibility—i.e., the acoustic-phonetic decoding of the message—to scalar ratings that investigate comprehensibility—the effort required to reconstruct the message (for a complete list, see Thomson 18-20). These assessment tools are rather simple and user-friendly, which makes them an attractive device that, in principle, anyone can use without much difficulty, even if not linguistically trained (Isbell 90; Saito et al., "Using Listener Judgments" 459).

However, human perceptions could be influenced by their personal beliefs, context and previous knowledge. For instance, listeners may have different expectations for L1 and L2 speakers or competence standards according to the students' educational, professional, or socio-economic background, which can influence how lenient they are when evaluating. Besides, the raters' familiarity with the students' L1 and the peculiarities of the community's L2 speech can also affect their judgments (Derwing and Munro 117; Teló et al. 1025).

The evaluation task and the previous training received can also influence the results obtained: on the one hand, the use of different scales or settings (online vs. offline) might lead to more or less accurate decisions (Derwing and Munro 117); on the other hand, a more limited or broader understanding of the concepts to assess before conducting the evaluation task can lead to different conclusions (O'Brien 738; Saito et al., "Using Listener Judgments" 456). These and other factors will be tackled in the forthcoming sections.

1.1. Coping with subjectivity

The use of human evaluators implies a certain degree of subjectivity in their judgments; thus, several techniques have been suggested to guarantee reliable outcomes. First, listeners may have a different idea of what *intelligibility* is about and focus on different pronunciation cues when rating. O'Brien (719) discussed the understanding of the dimension of *fluency*: while it could be understood from a narrow perspective as the 'fluidity of speech', it could also be interpreted in a broader sense as a synonym of proficiency (Lennon 389). Comprehensibility has also been considered a synonym of intelligibility when the latter is understood as general listening comprehension (Thomson 14). For this reason, it is crucial to provide clear instructions and a descriptive explanation of the items to evaluate before conducting the task. Additionally, initial training can also help listeners calibrate their judgments, even if they are working with inexperienced raters (Saito et al., "Using Listener Judgments" 458).

Second, listeners may be more or less familiar with the subjects' native language, which may lead to biased assumptions impaired by social prejudices or a previous awareness of the common mistakes that a specific group of speakers make (Derwing and Munro 117; Teló et al. 1025). Therefore, the knowledge and relationship the raters have with both the first (L1) and the second (L2) languages under study may influence whether they apply the scale in a more rigorous or flexible way (Isbell 107-108; Munro and Derwing, "Foreign Accent" 93-94; Munro and Derwing, "A Prospectus" 27-28; Southwood and Flege 336; Thomson 25). To further examine this issue, Teló et al. (1014) asked raters to comment on the criteria they had followed. In fact, gathering information about the listeners' beliefs and background can help better understand raters' leniency.

Third, other extrinsic agents such as fatigue, environmental noise, or family duties can also influence the listeners' opinions, since they act as distractions and alter the conditions under which the test is performed (Munro and Derwing, "A Prospectus" 27; Thomson 24). Thus, it is important to recruit enough judges and present samples in a randomized order to diminish the negative effects of these phenomena (Nagle, *A Guide* 143-149). Furthermore, it is strongly advisable to establish appropriate conditions to take the test and conduct statistical analyses of reliability to ensure the consistency of the judgments (Munro and Derwing, "A Prospectus" 30-31).

A key issue concerning listeners is how to recruit them. In English as a Foreign Language (EFL) contexts, where generally the target language is not spoken, it is difficult to obtain a sufficiently large number of listeners to provide reliable estimations. Native speakers of the target language tend to be scarce, and L2 speakers might not meet the established standards of the study: for instance, raters might be required to assess L2 speakers' comprehensibility, but results may vary if these speakers are teachers of the target language or not, due to their knowledge of both the first and the second languages. Nagle ("Developing" 10) introduced the possibility to use online tools such as Amazon's *Mechanical Turk* to crowdsource raters, an option that is worth exploring in detail, since it allows researchers to target a specific listener profile. Platforms like *Prolific* or *PsycoPY* are also frequently used with this purpose. However, this means of recruitment might lead to other problems, such as the raters' availability to take the test in person, since many might not live in town or in a nearby city. A potential solution to the problem is to request raters to participate remotely through online surveys; however, controlling some of the extrinsic agents mentioned before could become an arduous task.

1.2. The use of numeric scales

On the other hand, the test itself may undermine the listeners' appraisal as well. Numeric scales are among the most commonly employed tools to assess pronunciation, since they offer the possibility to quantify perception. However, their effectiveness has been the focus of extensive research and discussion.

A 9-point scale, favored by Derwing and Munro, has become the standard scale in comprehensibility studies (Thomson 19). Earlier, Southwood and Flege (346) had rejected the use of small numerical scales, such as 5- or 7-point scales, in an accentedness study, arguing that variance cannot be estimated above a certain level on these scales and may therefore lead to a ceiling effect. In contrast, Isaacs and Thomson (154) examined the raters' evaluation process in depth and found out that some raters considered a 9-point scale confusing due to the wide range of possibilities it offers, while they thought that the 5-point scale did not give rise to as many doubts. In a recent study, Nagle applied a many-facet Rasch analysis to examine the severity of raters' evaluating comprehensibility, fluency and accentedness via a 9-point scale. Based on the findings obtained, Nagle concluded that the scale steps were compressed for all three items tested, suggesting that a 5- or 7-point scale would have been more convenient due to the homogeneity of the speakers ("Developing" 19).

Other possibilities beyond the use of Likert scales have also been explored. For instance, Saito et al. ("Re-Examining" 145) effectively applied a 1,000-point continuous sliding scale for pronunciation assessment purposes. Therefore, using one scale or another could also be decisive to obtain trustworthy results.

1.3. Towards a mixed-method assessment

As previously discussed, thorough testing of the validity of rater-based assessment tools and their application help alleviate potential misconceptions and misguiding results (Harding 36-38), but full neutrality cannot be guaranteed. Ghanem and Kang (115) and Van Moere and Suzuki (149) pointed out that one of today's challenges in second language pronunciation research is to reconcile the communicative veracity obtained from listeners' judgments with the scientific objectivity of instrumental measures. Technological advances like speech recognition software are a potential solution, since they can offer more consistency, transparency, and accessibility to the assessment (see Babaeian 63-65 for a review). However, more research needs to be conducted to prove its reliability (Kang and Ginther 6). Meanwhile, a practical solution is to conduct studies based on a mixed-method assessment that uses already-tested instrumental tools, such as acoustic measurements, combined with listeners' perceptions, which have already helped to provide more evidence for detecting the linguistic aspects that influence comprehensibility and fluency (Derwing et al., "Second

Language Fluency" 655; Isaacs and Trofimovich 475-476; Kang et al. 554; Lennon 387; Nagle, *A Guide* 135-136; Riggenbach 423; Saito et al., "Re-Examining" 141).

1.4. Rhythm as a comprehensibility and fluency enhancement

Rhythm anticipates syntactic and lexical information and helps organize speech in a language (Derwing and Munro 60-61). In fact, rhythmic patterns evoke an acoustic illusion known as isochrony that might affect intelligibility, comprehensibility and fluency when communicating in a second language, both in terms of production and perception (Celce-Murcia et al. 131; Derwing and Munro 60-61; Levis 127-128). Hence, if the L1 and the L2 differ considerably as far as rhythm is concerned, misunderstandings and communication breakdowns tend to arise. L2 learners are not usually conscious about the rhythmic differences between their mother tongue and the target language; consequently, it is difficult for them to correct their own speech and try not to transfer the L1 rhythm when speaking in the L2.

Spanish/Catalan natives studying English often face this issue. While stress plays a crucial role in the length of syllables in English (among other factors), it does not in Spanish or Catalan, whose syllables last approximately the same regardless of whether these are stressed or unstressed (for a detailed description of different language rhythms, see Dauer 51-52). A pronunciation teaching study that has aimed at examining the effectiveness of teaching rhythm to Spanish EFL learners is Chela-Flores' (122). The experiment consisted in teaching word-decontextualized rhythmic patterns to tertiary students in Venezuela for a semester. Apparently, subjects perceived and recognized words with the taught rhythmic patterns considerably better after instruction, but they were tested just under controlled circumstances. Besides, neither the assessment process nor the tools used were described.

Differences in rhythm have been acoustically measured by means of different rhythmic metrics. The use of one rhythmic measurement or another will often depend on the rhythmic feature to analyze. As previously mentioned, duration is a parameter that distinguishes Spanish/Catalan from English. After investigating the metrics that could better capture the rhythm acquisition of L2 speakers of Spanish and English, White and Mattys ("Calibrating Rhythm" 520; "Rhythmic Typology" 253) concluded that the proportion of vocalic interval durations in a given sentence (%V) and the variation coefficient indices for vowels (VarcoV) were the most reliable, both related to syllable length. Based on this assumption, the project presented hereafter performed acoustic analysis with the two measurements.

1.5. The project: Previous findings

The main goal of the project is to examine the effectiveness of explicit rhythm instruction to improve Spanish/Catalan ESP students' comprehensibility and fluency. Participants were divided into an experimental group, which received explicit rhythm instruction, and a control group, which did not (for a detailed description of the treatment, see the following section). Previous studies associated with the project aimed at investigating the L1 negative transfer after instruction first (Quesada Vázquez, "Pronunciation Instruction" 163; "Rhythm Instruction to Enhance ESP Students' Prosody", forthcoming; Quesada Vázquez and Romero, "The Improvement" 104; "ESP within ESP" 209). To this end, measurements of VarcoV and %V were obtained. Broadly speaking, it was concluded that, despite a lack of definitive statistical significance, students who received explicit rhythm instruction exhibited higher VarcoV values and lower %V values after treatment, which reflects a more target-like rhythm and, hence, a reduction in their L1 negative transfer. In fact, a significant effect size was revealed when VarcoV values were considered, which supports the impact of explicit rhythm instruction in the students' performances.

Additionally, pausing was also investigated (Quesada Vázquez, "Rhythm Instruction to Improve ESP Students' Fluency" 63), as it is also thought to influence L2 prosody, especially when fluency is considered (Lennon 413-414; Riggenbach 438; Wood 1). To this end, the total number of pauses and the number of unfilled pauses students made were examined. Results showed that both groups decreased the number of pauses (both the total number and the unfilled ones), but the experimental group made fewer after instruction. Besides, the statistical analysis conducted based on ANOVAs showed statistical significance for time and group, but not for their interaction.

This paper attempts to examine whether explicit rhythm instruction is also beneficial for students' comprehensibility and fluency. In this line, the following research questions are formulated:

RQ1: To what extent does explicit rhythm instruction influence students' comprehensibility and fluency?
RQ2: To what extent do objective (acoustic analysis) and subjective measurements (listeners' ratings) correlate?

2. Methodology

A classroom-based study (Derwing and Munro 84) was designed to test the effectiveness of explicit rhythm instruction within the ESP classroom to improve students' comprehensibility and fluency. A longitudinal experiment was conducted: a ten-week pronunciation module was created and embedded within a compulsory technical English course taught from February to May 2017 to first-year engineering undergraduates at Rovira i Virgili University (Tarragona, Spain). Although all the participants took pronunciation instruction, some of them took explicit rhythm training (experimental group) while the others did not (control group).

The experiment followed a pretest-posttest design to assess the progress of the subjects. All the students were recorded individually before and after instruction in three isolated rooms at the university library with two Sony PCM-M1O and a Zoom H4nsp recorders. For both the pretest and the posttest, students were requested to read ten sentences and a text aloud, introduce themselves, and give their opinion on social networking. While previous research has focused only on the acoustic analysis of read aloud sentences, the present study examines students' extemporaneous speeches, i.e. their opinions.

2.1. Participants

The pronunciation module was designed considering the 298 students enrolled for the course. However, only those who attended 90% of the sessions were taken as final participants of the study due to its longitudinal nature. Unfortunately, the number of participants decreased dramatically owing to absenteeism, so only 42 students (21 experimental; 21 control) were the final subjects.

The "introduce yourself" exercise of the assessment test was used to gather information about the participants' age and linguistic background. All the students in the experimental group were 18-year-old first-year undergraduates who accessed university from high school, while there were five students in the control group who were over eighteen (two 19-year-olds, two 20-year-olds, and one 21-year-old). One of the 19-year-old students and the two 20-year-old students were taking the course for the second or third time. Most participants were balanced bilinguals of Spanish and Catalan coming from different cities in Catalonia. One student from the experimental group was from Venezuela and another from the control group was from Madrid, both having Spanish as their dominant language. Additionally, four students reported a different mother tongue: two Arabic, one Portuguese, and one Bulgarian. However, all of them had pursued primary and secondary education and taken the university access examination in Catalonia; thus, their command of both Spanish and Catalan was high.

The placement test that students had to take at the beginning of the course as part of its regular evaluation system shed light on their initial English level. Students exhibited a broad range of English proficiency levels within the groups: for the experimental group, ten students had a pre-intermediate level, six an intermediate level, and five an advanced level; for the control group, four students had a pre-intermediate level, eleven had an intermediate level, and six had an advanced level.

2.2. Procedure

For ten weeks, all the students received pronunciation training for thirty minutes a week. Instruction was "scaffolded": the first sessions dealt with features influencing the pronunciation of the basic unit of speech (the word); then, the following sessions covered the pronunciation of longer utterances (the sentence); and finally, the last sessions focused on more complex units (the discourse).

On the other hand, each individual session followed Celce-Murcia et al.'s (36) steps aimed at teaching communicatively: The first part of each session consisted of a description and analysis of the item to work on. The teacher's explanations were complemented with videos, images, and kinesthetic activities; the second part was based on a perception activity in which students were encouraged to identify the target item; the third and final part of the session consisted of a series of production exercises, starting with controlled/guided training and finishing with communicative practice. This approach was used in both experimental and control sessions; the only difference in treatment was thus the explicit explanation and practice of rhythm instruction.

The materials and exercises used for the instruction, available on Mora-Plaza et al's *Tools* repository, were especially created for this experiment using online resources (videos, images, etc.), or adapted from pronunciation books and research papers, modifying rhythm activities to use technical vocabulary. These changes attempted to encourage students' motivation (Anderson-Hsieh 201) while adapting the content of the module to the students' needs (Dudley-Evans and St John 3). Moreover, the grammar and lexicon used in the sessions had already been taught in class; hence, students were familiar with the content, which facilitated the learning process.

2.3. Data analysis

An online Google form was designed to record students' answers to section #4 of the pronunciation test (i.e., students' opinions on social media)—forty-two for the control group and forty-two for the experimental group—before and after instruction, to assess students' progress on comprehensibility and fluency. An individual audio file was created for each of the opinions. By using *Audacity* 2.2.2, audios were normalized at -1 dB. However, a few audios were still too quiet and required further volume adjustment to ensure a comfortable listening level without distortions. On the other hand, files were cut into forty-five seconds when the contributions were too long (Derwing et al., "Evidence" 404). When possible, the beginning of the opinion segment was selected for assessment; otherwise, the part with fewer hesitations and pauses was used. The end point was established at a natural pause in the argumentation.

The beginning of the form featured instructions for the listeners on how to carry out the test and explanations of the concepts to analyze, based on Derwing and Munro's (76) and Wood's (9) definitions of comprehensibility and fluency, respectively: *comprehensibility* was defined as how easy the listener found it to understand the speech, while *fluency* was operationalized as the extent to which the speaker delivered the speech in a fluid way, without hesitations, pauses, false starts, repetitions or fragmentations.

Seven native speakers of English (two men and five women) were recruited to judge the students' extemporaneous opinions. They came from different English-speaking countries: four from the United Kingdom, one from Canada, one from the United States of America and one from Singapore. Listeners were also requested to provide their age and nationality, the years of residence in Catalonia, and the years of teaching English as a foreign language. They were all considered skilled listeners due to their experience in teaching English as a foreign language to Spanish/Catalan speakers in Catalonia. Nevertheless, the length of teaching experience varied considerably, between one and thirty-five years (M = 17.43; SD = 14.8). All of them were living in Catalonia at the time they took the test, so they were familiar with the Catalan and Spanish accents to some extent, although their length of residence varied greatly, from moths to years (M = 15.86; SD = 15.32). They all participated voluntarily in the project.

The questionnaire was divided into seven sections, each of which contained twelve audios. All the audios, both pre- and post-treatment, were randomly distributed within the form using *Research Randomizer*; however, the same order was presented to all listeners. For each file, listeners opened a link where they listened to the audio and then had to decide how comprehensible and fluent that student was by using a 5-point Likert scale, ranging from 1—"not at all"—to 5—"totally". A short scale was considered suitable to ensure user-friendliness (Isaacs and Thomson 154; Nagle, "Developing" 19). Raters listened to the recording once and were asked to assess it twice, once for each dimension (comprehensibility and fluency).

Listeners had to take the test online due to the inability to bring all of them together in the same place on the same day. The limitations of the online form, which did not allow to save the answers to take a break, forced listeners to take the test in one sitting. The test lasted approximately one hour and a half. Therefore, the fatigue factor could not be controlled for and will be considered when analyzing and discussing the results.

2.4. Statistical analysis

Several analyses were performed in order to test the effectiveness of the instruction to improve students' comprehensibility and fluency. First, intra-class correlation coefficient ICC tests were run on the listeners' scores to guarantee the reliability of the results. Second, Friedman tests were conducted to study the general impact of the instruction, as not all the data obtained from the ratings were normally distributed (see following section). Lastly, Spearman's and Kendall's Tau b coefficients were calculated to further examine the relationship between both comprehensibility and fluency and previous acoustic analysis. For all the tests, the alpha value was set at .05.

3. Results

An intraclass correlation coefficient (ICC) test was computed to examine the reliability of the raters for comprehensibility and fluency separately. A two-way random effect model based on mean ratings and consistency assessed the uniformity of the scoring. Results reported a good mean estimation (α = .873) along with a good-to-excellent 95% confidence interval (IC = .826-.910) for comprehensibility scores, and a good mean estimation (α = .895.) along with a good-to-excellent 95% confidence interval (IC = .857-.926) for fluency scores. Hence, scores were considered reliable.

Table 1 illustrates the means of the raters' scores for both comprehensibility and fluency. Marks were distributed according to time (T1 = pretest; T2 = posttest) and group (experimental; control). Regarding comprehensibility, the control group was considered to improve slightly, while the experimental group showed certain signs of deterioration. Therefore, the comparison of the means suggests that the control group was more comprehensible after instruction than the experimental group, even though scores are approximately equal, oscillating between 3 and 3.3 scores.

Table 1. Raters' mean scores (M) and standard deviations (SD) for comprehensibility and fluency ratings

	Control group				**Experimental group**			
	T1		**T2**		**T1**		**T2**	
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Comprehensibility	3.05	0.74	3.11	0.71	3.29	0.63	3.27	0.66
Fluency	2.81	0.77	2.86	0.72	2.98	0.59	3.01	0.70

As for the fluency ratings, both groups tend to improve after treatment: the control group appears to perform better than the experimental group after instruction, though the difference is very small.

To further investigate the students' comprehensibility and fluency progress according to the type of treatment received, four Friedman tests were performed (see Table 2). As mentioned in the previous section, data were not normally distributed for all the levels of the variables: for comprehensibility scores, the experimental posttest variable does not show a normal distribution (W = .884 , p = .018); for fluency scores, both the experimental pretest (W = .904, p = .041) and posttest (W = .902 , p = .038) variables do not show a normal distribution. None of the tests run shows statistical significance.

Table 2. Friedman tests of comprehensibility and fluency scores

	Group	**X^2**	***df***	***p*-value**	**Kendall's W**
Comprehensibility	Control	.429	1	.513	.020
	Experimental	.059	1	.808	.003
Fluency	Control	.800	1	.371	.038
	Experimental	1.471	1	.225	.070

Spearman's and Kendall's Tau b coefficients were also calculated to investigate the extent to which comprehensibility and fluency results correlated. As displayed in Table 3, all four tests show statistical significance. Hence, the tests suggest that the students' level of comprehensibility is related to their level of fluency.

Table 3. Comprehensibility – fluency Pearson Spearman's and Kendall's Tau b correlation coefficients

		Spearman's		**Kendall's**	
Group	**Time**	**rho**	***p***	**Tau b**	***p***
Control	1	.891	.01**	.759	.01**
	2	.893	.01**	.770	.01**
Experimental	1	.710	.01**	.588	.01**
	2	.755	.01**	.625	.01**

Note. * $p < .05$, ** $p < .01$.

Correlations between ratings and the different acoustic measurements conducted (VarcoV, %V, total number of pauses and unfilled pauses) were also analyzed. As shown in Table 4, correlations between comprehensibility scores and VarcoV values indicate statistical significance for the experimental group in the pretest [$r_s(19) = .537$, $p = .012$; $\tau b = .361$, $p = .026$]. Regarding correlations between fluency scores and VarcoV values, no statistical significance is found (see Table 5).

Table 4. Comprehensibility – VarcoV Pearson Spearman's and Kendall's Tau b correlation coefficients

		Spearman's		**Kendall's**	
Group	**Time**	**rho**	***p***	**Tau b**	***p***
Control	1	-.047	.840	.015	.927
	2	.419	.059	.300	.064
Experimental	1	.537	.012*	.361	.026*
	2	.350	.120	.194	.234

Note. * $p < .05$, ** $p < .01$.

Table 5. Fluency – VarcoV Pearson Spearman's and Kendall's Tau b correlation coefficients

		Spearman's		**Kendall's**	
Group	**Time**	**rho**	***p***	**Tau b**	***p***
Control	1	-.049	.834	-.014	.928
	2	.273	.231	.203	.212
Experimental	1	.189	.411	.132	.422
	2	.206	.371	.129	.427

Note. * $p < .05$, ** $p < .01$.

Interestingly, the control group reveals a negative correlation in the pretest for the relationship of VarcoV values with comprehensibility [r_s (19) = -.047, p = .0840] and fluency [r_s (19) = -.049, p = .834; τb = -.014, p = .928], respectively. Both VarcoV values and scores are supposed to increase if an improvement is detected; thus, a negative correlation might suggest that the lower the students' VarcoV values, the better their comprehensibility and fluency, challenging the idea that better rhythm adaptation leads to better prosody (for a full discussion, see the forthcoming section).

Correlations were also performed for %V values and comprehensibility and fluency ratings, respectively (see Tables 6 and 7). Only the correlation between comprehensibility and %V for the control group in the pretest displays statistical significance [r_s (19) = -.467, p = .033; τb = -.285, p = .017]. %V is thought to decrease if students adopt a more English-like rhythm; hence, a negative correlation between %V and comprehensibility and fluency scores was expected. By contrast, the correlation between %V and comprehensibility scores for the experimental group after treatment is positive [r_s (19) = .030, p = .897; τb = .025, p = .879].

Table 6. Comprehensibility – %V Pearson Spearman's and Kendall's Tau b correlation coefficients

		Spearman's		Kendall's	
Group	**Time**	**rho**	***p***	**Tau b**	***p***
Control	1	-.467	.033*	-.285	.017*
	2	-.067	.772	-.054	.738
Experimental	1	-.196	.393	-.153	.344
	2	.030	.897	.025	.879

Note. * $p < .05$, ** $p < .01$.

Table 7. Fluency – %V Pearson Spearman's and Kendall's Tau b correlation coefficients

		Spearman's		Kendall's	
Group	**Time**	**rho**	***p***	**Tau b**	***p***
Control	1	-.293	.198	-.246	.123
	2	-.005	.984	-.020	.903
Experimental	1	-.393	.078	-.233	.156
	2	-.072	.756	-.050	.760

Note. * $p < .05$, ** $p < .01$.

Finally, correlations between pausing and comprehensibility and fluency ratings were also examined. Correlation tests were first run to investigate the relationship between the total number of pauses and comprehensibility and fluency scores, respectively. Results show that the total number of pauses and comprehensibility ratings are significantly correlated for the control group before [r_s (19) = -.595, p = .004; τb = -.406, p = .012] and after [r_s (19) = -.613, p = .003; τb = -.442, p = .007] treatment, but not for the experimental group (see Table 8).

Table 8. Fluency – total number of pauses Pearson Spearman's and Kendall's Tau b correlation coefficients

		Spearman's		**Kendall's**	
Group	**Time**	**rho**	***p***	**Tau b**	***p***
Control	1	-.595	.004**	-.406	.012*
	2	-.613	.003**	-.442	.007**
Experimental	1	-.089	.703	-.045	.783
	2	.198	.389	.125	.444

Note. * $p < .05$, ** $p < .01$.

The total number of pauses and fluency are also significantly correlated for the control group in Time 1 [r_s (19) = -.575, p = .006; τb = -.402, p = .012] and Time 2 [r_s (19) = -.494, p = .023; τb = -.334, p = .041], while the experimental group did not reveal statistical significance (see Table 9). Again, all the tests show a negative correlation except for the experimental group after treatment for both comprehensibility and fluency.

Table 9. Fluency – total number of pauses Pearson Spearman's and Kendall's Tau b correlation coefficients

		Spearman's		**Kendall's**	
Group	**Time**	**rho**	***p***	**Tau b**	***p***
Control	1	-.575	.006**	-.402	.012*
	2	-.494	.023*	-.334	.041*
Experimental	1	-.228	.321	-.179	.279
	2	.065	.781	.050	.760

Note. * $p < .05$, ** $p < .01$.

A similar behavior is observed when correlations between unfilled pauses and comprehensibility and fluency ratings were run (see Tables 10 and 11): while the control group exhibits statistical significance for both dimensions in Time 1 [comprehensibility: r_s (19) = -.671, p = <.001; τb = -.487, p = .003; fluency: r_s (19) = -.629, p = .002; τb = -.433, p = .007] and Time 2 [comprehensibility: r_s (19) = -.636, p = .002;

τb = -.474, p = .004; fluency: r_s (19) = -.496, p = .022; τb = -.346, p = .035], the experimental group does not indicate statistical significance. Besides, the experimental group reveals positive correlations for both comprehensibility and fluency after treatment.

Table 10. Comprehensibility – unfilled pauses Pearson Spearman's and Kendall's Tau b correlation coefficients

		Spearman's		**Kendall's**	
Group	**Time**	**rho**	***p***	**Tau b**	***p***
Control	1	-.671	<.001**	-.487	.003**
	2	-.636	.002**	-.474	.004**
Experimental	1	-.145	.530	-.106	.520
	2	.231	.314	.140	.392

Note. * $p < .05$, ** $p < .01$.

Table 11. Fluency – unfilled pauses Pearson Spearman's and Kendall's Tau b correlation coefficients

		Spearman's		**Kendall's**	
Group	**Time**	**rho**	***p***	**Tau b**	***p***
Control	1	-.629	.002**	.433	.007**
	2	-.496	.022*	-.346	.035*
Experimental	1	-.243	.288	-.191	.252
	2	.045	.847	.030	.854

Note. * $p < .05$, ** $p < .01$.

4. Discussion and conclusion

The descriptive analysis of the comprehensibility scores showed that the overall performance of the experimental group slightly worsened after instruction, although both groups—experimental and control—were observed to remain almost stable. Regarding fluency, both groups displayed improvement after training: the control group appeared to improve somewhat more, but again changes were minimal for both groups. In addition, Friedman tests revealed no statistical significance, while both dimensions—comprehensibility and fluency—showed high correlations for all the levels of the variables under study. Hence, these findings suggest that explicit rhythm instruction did not enhance students' comprehensibility and fluency (RQ1).

Nevertheless, previous studies related to the project showed improvement for the experimental group when acoustic measurements were applied: overall, learners appeared to increase their VarcoV values and decrease their %V values, showing a better adaptation to a more stressed-timed rhythm (Quesada Vázquez, "Pronunciation instruction" 163; "Rhythm Instruction to Enhance ESP Students' Prosody," forthcoming; Quesada Vázquez and Romero, "The Improvement" 104; "ESP within ESP" 209). Besides, they were also observed to improve their pausing, which should lead to more fluent speech (Quesada Vázquez, "Rhythm Instruction to Improve ESP Students' Fluency" 63).

Concerning the correlations between subjective measurements and acoustic analyses (RQ2), different observations could be made: As for the relationship between comprehensibility and fluency and VarcoV, only comprehensibility and VarcoV values for the experimental group in Time 1 showed statistical significance. Additionally, VarcoV values and comprehensibility and fluency, respectively, were negatively correlated. These findings challenge the belief that a better adoption of a more target-like rhythm will boost students' comprehensibility and fluency, as suggested by several researchers in the field (Celce-Murcia et al. 131; Derwing and Munro 60-61; Levis 127-128). However, a different behavior was observed when considering %V correlations: in this case, the control group in Time 1 showed statistical significance when the relationship between comprehensibility and %V was studied, whereas it was the experimental group in Time 2 that revealed a positive correlation.

Therefore, VarcoV and %V behaved differently. White and Mattys ("Calibrating Rhythm" 520; "Rhythmic Typology" 253) considered both metrics reliable, which is why both were analyzed in this project. However, Quesada Vázquez ("Rhythm Instruction to Enhance ESP Students' Prosody," forthcoming) observed that they do not correlate and that VarcoV values tend to yield stronger results. Further research should be conducted to reach more robust conclusions. Nonetheless, both rhythmic metrics showed signs of improvement, while subjective ratings did not. Perhaps the length of the treatment was not sufficient to produce large-scale changes, but it begun to take effect on a smaller scale: an adoption of a more target-like rhythm indicates that explicit instruction can be beneficial, though students might need more practice to show improvement in their overall prosody (Quesada Vázquez, "Rhythm Instruction to Enhance ESP Students' Prosody," forthcoming).

Regarding pausing, significant results were found for both the total number of pauses and unfilled pauses when correlated with both comprehensibility and fluency results for the control group before and after treatment, but not for the experimental group. Besides, the experimental group showed positive correlations after treatment for both dimensions. These findings, together with the positive correlation found when %V was examined, appear to suggest that there might have been some difficulties with the subjective measurements, particularly when the performances of the experimental group were rated. A potential explanation is the existence of a ceiling effect due to the choice of experienced raters, since when examining individual ratings closely, experimental low-level students obtained rather high scores for their pretest performances (see Quesada Vázquez, *The Introduction of Rhythm* 142 for further information). Listeners were ESL teachers that had lived in Catalonia and were therefore familiar with the Spanish/Catalan language and the most common mistakes the speakers of those languages make when speaking English. Although the instructions specified that they should rate as natives and not as second language teachers, the familiarity with the language may have eased the degree to which they understood the speech and, thus, their criterion could have become more lenient (Harding 38; Isaacs and Thomson 139). Southwood and Flege (347) brought up this possibility when analyzing accentedness of Italian speakers of English. This fact, hence, could explain why low-level students obtained higher scores than expected.

The type of scale chosen could also have influenced ratings. For this study, a 5-point rather than the standardized 9-point scale was preferred because it was expected to be more user-friendly, as it simplifies the scope, leaving less room for doubts and hesitations. However, Southwood and Flege (346) had already warned about the possibility of a ceiling effect when a short scale like a 7-point was used, recommending the use of longer scales. Therefore, the choice of scale could partly explain the existence of a ceiling effect as well.

Indeed, the test suffered from different limitations that could have influenced the results. It should also be borne in mind that the test was conducted remotely. The main drawback of this type of assessment test is the impossibility of controlling the context in which it is performed: for example, the quality of the audio may not have been appropriate due to the conditions of the equipment, or external factors such as the presence of family members or other duties could have interfered in the performance. Munro and Derwing (25) already warned about the drawbacks of listeners carrying out assessment tests at home. Ideally, it would have been more convenient to bring all the raters together in the same space and have them complete the test under the same conditions. However, incompatibility in the raters' schedules made this unfeasible, making the design of an online form the only viable alternative. Other researchers in the field (Kang et al. 557) have used an online test before, but little was reported about its performance constraints. They did specify, though, that they used a trial sample that was not included in the analysis to create a reference that guided the raters' criterion.

Other researchers also used checks by embedding native samples (Derwing et al., "Evidence" 400; Derwing et al., "Second Language Fluency" 663; Isaacs and Thomson 141). Trial or native samples to control the listeners' partiality were not included in this study's test, because raters were both native speakers and EFL teachers; hence, they were considered to have sufficient training to apply the right criterion. However, it is undeniable that providing a native sample could have helped prevent overrating: due to the disparity of levels within the groups, it would be easy to consider an average speaker as very fluent and comprehensible after listening to a low-level student who is clearly struggling. Therefore, a native speech sample would have acted as a benchmark of what constitutes a truly strong performance, recalculating the perception of learners' speeches. Consequently, the choice of not using native samples in the test could explain why some speeches were overrated, thereby causing the ceiling effect phenomenon.

Although the limitations of an online form affected the conditions of the performance, there are further factors that should also be taken into consideration. As previously mentioned, fatigue could also have played an important role in raters' judgments. The total duration was calculated at one and a half hours due to the large number of recordings under assessment and, since the form did not allow to save scores during the process, raters had to listen to all the recordings in one sitting. Munro and Derwing (27) pointed out the threat that fatigue could pose in rating tasks and recommended that assessment sessions should not exceed one hour, with appropriate breaks in between. Therefore, it is highly possible that fatigue influenced the scoring (Thomson 24). Again, an in-person assessment would have allowed to divide the task into different sessions and, hence, alleviate fatigue (Derwing et al., "Evidence" 401; Munro and Derwing, "A Prospectus" 29; Saito et al., "Second Language Speech Production" 224), but the circumstances of the experiment made it unfeasible.

Despite a thorough examination of the available options and the potential variables influencing the rating test before conducting it, factors such as the online nature of the form, the type of listeners, fatigue, and the chosen scale are assessment issues that may have affected the performance of the rating test. For this reason, if the project had just relied on raters' judgments, results would have completely failed to support an improvement in students' performances, especially when rhythm was explicitly taught. However, further examination via acoustic measurements has allowed not only to detect problems with the rating test, but also to discern certain tendencies that support a positive effectiveness of the pronunciation module.

Last, but not least, results revealed a significant correlation between comprehensibility and fluency scores, supporting Levis' (27) claims on their tight relationship, in that disfluent speech can break the native listener's train of thought when interpreting the message and, as a consequence, hinder comprehension.

5. Future research

After examining the various limitations that human ratings can pose when used as an assessment tool in pronunciation teaching research, the use of a mixed-method approach is strongly encouraged (Nagle, *A Guide* 135-136; Teló et al. 1025). This could involve different types of quantitative data (i.e., rater judgments and acoustic measurements, for instance) or combining both quantitative and qualitative data (e.g., rater judgments and students'/teachers' questionnaires and/or interviews). This type of assessment will limit the drawbacks of perception tests and, while providing a wider picture of the learning process, lead to more conclusive and reliable findings.

Acknowledgments

This research was funded by project PID2023-150120NB-I00 from the Spanish Ministry of Economy and Business.

Works cited

Anderson-Hsieh, Janet. "Teaching Suprasegmentals to International Teaching Assistants Using Field-Specific Materials." *English for Specific Purposes*, vol. 9, 1990, pp. 195-214. https://doi.org/10.1016/0889-4906(90)90013-3.

Audacity. Version 2.2.2, Muse Group & Contributors, 20 Feb. 2018.

Babaeian, Ali. "Pronunciation Assessment: Traditional vs Modern Modes." *Journal of Studies in Education*, vol. 13, no. 4, 2023, pp. 89-97. https://doi.org/10.5296/jse.v13i4.21417.

Celce-Murcia, Marian, et al. *Teaching Pronunciation: A Reference for Teachers of English to Speakers of Other Languages*, Cambrige UP, 1996.

Chela-Flores, Bertha. "Rhythmic Patterns as Basic Units in Pronunciation Teaching." *Onomázein*, vol. 2, 1997, pp. 111-134.

Dauer, Rebecca M. "Stress-Timing and Syllable-timing Reanalyzed." *Journal of Phonetics*, vol. 11, no. 52, 1983, pp. 51-62.

Derwing, Tracey M., and Murray J. Munro. *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research*, John Benjamins, 2015.

Derwing, Tracey M., et al. "Evidence in Favor of a Broad Framework for Pronunciation Instruction." *Language Learning*, vol. 48, no. 3, 1998, pp. 393-410. https://doi.org/10. 1111/0023-8333.00047.

Derwing, Tracey M., et al. "Second Language Fluency: Judgments on Different Tasks." *Language Learning*, vol. 54, no. 4, 2004, pp. 655-679.

Dudley-Evans, Tony, and Maggie Jo St John. *Developments in English for Specific Purposes. A Multi-Disciplinary Approach*, Cambridge UP, 1998.

Ghanem, Romy, and Okim Kang. "Pronunciation Features in Rating Criteria." *Assessment in Second Language Pronunciation*, edited by Okim Kang and April Ginther, Routledge, 2018, pp. 115-136.

Harding, Luke. "Validity in Pronunciation Assessment." *Assessment in Second Language Pronunciation*, edited by Okim Kang and April Ginther, Routledge, 2018, pp. 30-48.

Isaacs, Talia, and Ron I. Thomson. "Rater Experience, Rating Scale Length, and Judgments of L2 Pronunciation: Revisiting Research Conventions." *Language Assessment Quarterly*, vol. 10, 2013, pp. 135-159. https://doi.org/10.1080/15434303.2013.769545.

Isaacs, Talia, and Pavel Trofimovich. "Deconstructing Comprehensibility: Identifying the Linguistic Influences on Listeners' L2 Comprehensibility Ratings." *Studies in Second Language Acquisition*, vol. 34, 2012, pp. 475-505. https://doi.org/10.1017/ S0272263112000150.

Isbell, Dan R. "Assessing Pronunciation for Research Purposes with Listener-Based Numerical Scales." *Assessment in Second Language Pronunciation*, edited by Okim Kang and April Ginther, Routledge, 2018, pp. 89-112.

Kang, Okim, and April Ginther, editors. *Assessment in Second Language Pronunciation*, Routledge, 2018.

Kang, Okim, et al. "Suprasegmental Measures of Accentedness and Judgments of Language Learner Proficiency in Oral English." *The Modern Language Journal*, vol. 94, no. 4, 2010, pp. 554-566. https://doi.org/10.1111/j.1540-4781.2010.01091.x.

Lennon, Paul. "Investigating Fluency in EFL: A Quantitative Approach." *Language Learning*, vol. 40, no. 3, 1990, pp. 387-417. https://doi.org/10.1111/j.1467-1770.1990.tb00 669.x.

Levis, John M. *Intelligibility, Oral Communication, and the Teaching of Pronunciation*, Cambrige UP, 2018.

Mora-Plaza, Ingrid, et al. *Tools for Second Language Speech Research and Teaching*. 2022. http://sla-speech-tools.com.

Munro, Murray J., and Tracey M. Derwing. "Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second

Language Learners." *Language Learning*, vol. 45, no. 1, 1995, pp. 73-97. https://doi.org/10.1111/j.1467-1770. 1995.tb00963.x.

— "A Prospectus for Pronunciation Research in the 21st Century: A Point of View." *Journal of Second Language Pronunciation*, vol. 1, no. 1, 2015, pp. 11-42. https://doi.org/10.1075/jslp.1.1.01mun.

Nagle, Charles. *A Guide to Quantitative Research Methods in Second Language Pronunciation*, Routledge, 2025.

— "Developing and Validating a Methodology for Crowdsourcing L2 Speech Ratings in Amazon Mechanical Turk." *Journal of Second Language Pronunciation*, vol. 5, no. 2, 2019, pp. 1-30. https://doi.org/10.1075/jslp.18016.nag.

O'Brien, Mary G. "L2 Learners' Assessments of Accentedness, Fluency, and Comprehensibility of Native and Nonnative German Speech." *Language Learning*, vol. 64, no. 4, 2014, pp. 715-748. https://doi.org/10.1111/lang.12082.

Quesada Vázquez, Leticia. "Pronunciation Instruction in ESP Teaching to Enhance Students' Prosody." *ESP Teaching and Teacher Education: Current Theories and Practices*, edited by Salomi Papadima-Sophocleous, Elis Kakoulli Constantinou, and Christina Nicole Giannikas, 2019, pp. 163-176. Research-publishing.net. https://doi.org/10.14705/rpnet.2019.33.932.

— *The Introduction of Rhythm Instruction in the English as a Foreign Language Classroom to Improve the Comprehensibility and Fluency of English for Specific Purposes Students*. Honors thesis, Rovira i Virgili University, 2019.

— "Rhythm Instruction to Improve ESP Students' Fluency: A Study of Pausing." *Revista de Estudos Linguísticos da Universidade do Porto*, vol. 17, 2022, pp. 63-93. https://doi.org/10.21747/16466195/ling17a3.

— "Rhythm Instruction to Enhance ESP Students' Prosody: An Acoustic Study. *Proc. Forum Acousticum Euronoise*, forthcoming.

Quesada Vázquez, Leticia, and Joaquín Romero. "The Improvement of Spanish/Catalan EFL Students' Prosody by Means of Explicit Rhythm Instruction." *Proc. International Symposium on Applied Phonetics*, 2018, pp. 104-109. https://doi.org/10.21437/ISAPh.2018-19.

—"ESP Within ESP: The Design and Implementation of a Pronunciation Module in a Technical English Course". *Onomázein*, n.º NE VI, 2020, pp. 209-228. https://doi.org/10.7764/onomazein.ne6.11.

Riggenbach, Heidi. "Toward an Understanding of Fluency: A Microanalysis of Nonnative Speaker Conversations." *Discourse Processes*, vol. 14, no. 4, 1991, pp. 423-441. https://doi.org/ 10.1080/01638539109544795.

Saito, Kazuya, et al. "Using Listener Judgments to Investigate Linguistic Influences on L2 Comprehensibility and Accentedness: A Validation and Generalization Study." *Applied Linguistics*, vol. 38, no. 4, 2017, pp. 439-462. https://doi.org/10.1093/ applin/ amv047.

— "Second Language Speech Production: Investigating Linguistic Correlates of Comprehensibility and Accentedness for Learners at Different Ability Levels." *Applied Psycholinguistics,* vol. 37, 2016, pp. 217-240. https://doi.org/10.1017 /S0142716414000502.

Saito, Kazuya, et al. "Re-Examining Phonological and Lexical Correlates of Second Language Comprehensibility: The Role of Rater Experience." *Second Language Pronunciation Assessment*, edited by Talia Isaacs and Pavel Trofimovich, Multilingual Matters, 2017, pp. 259-271. https://doi.org/10.21832/ISAACS6848.

Southwood, M. Helen, and James E. Flege. "Scaling Foreign Accent: Direct Magnitude Estimation Versus Interval Scaling." *Clinical Linguistics & Phonetics*, vol. 13, no. 5, 1999, pp. 335-349.

Teló, Cesar, et al. "Accent Bias in Professional Evaluations: A Conceptual Replication Study in Brazil." *Applied Linguistics*, vol. 45, no. 6, 2024, pp. 1010-1029. https://doi.org/10.1093/applin/amae042.

Thomson, Ron I. "Measurement of Accentedness, Intelligibility, and Comprehensibility." *Assessment in Second Language Pronunciation*, Routledge, 2018, pp. 11-29.

Urbaniak, Geoffrey C., and Scott Plous. *Research Randomizer*, 1997. https://www.randomizer.org/.

Van Moere, Alistair, and Masanori Suzuki. "Using Speech Processing Technology in Assessing Pronunciation." *Assessment in Second Language Pronunciation*, edited by Okim Kang and April Ginther, Routledge, 2018, pp. 137-152.

White, Laurence, and Sven L. Mattys. "Calibrating Rhythm: First Language and Second Language Studies." *Journal of Phonetics*, vol. 35, 2007, pp. 501-522.

— "Rhythmic Typology and Variation in First and Second Languages." *Segmental and Prosodic Issues in Romance Phonology (Current Issues* *in Linguistic Theory Series)*, edited by Pilar Prieto, Joan Mascaró, and Maria Josep Solé, John Benjamins, 2007, pp. 237-257.

Wood, David. *Formulaic Language and Second Language Speech Fluency: Background, Evidence and Classroom Applications*, Continuum, 2010.

Received / Recibido: 13/3/2025 *Accepted / Aceptado: 7/7/2025*