International Journal of Applied Linguistics & English Literature
E-ISSN: 2200-3452 & P-ISSN: 2200-3592
www.ijalel.aiac.org.au

The Self, the Peer and the Teacher in the EFL Pronunciation Class: A Comparative Study on Assessment, Perceptions and Systematicity

Nuria Edo-Marzá*
Faculty of Humanities and Social Sciences, Department of English Studies, Universitat Jaume I , Avda. Sos Baynat s/n, 12071, Castellón, Spain

Corresponding Author: Nuria Edo-Marzá, E-mail: nedo@ang.uji.es

ARTICLE INFO

Article history
Received: July 04, 2017
Accepted: August 16, 2017
Published: December 01, 2017
Volume: 6 Issue: 7
Special Issue on Language & Literature
Advance access: September 2017

Conflicts of interest: None
Funding: Education and Innovation Research Project [2779/13]

ABSTRACT

This pilot study is aimed at describing, analysing and comparing self, peer and teacher quantitative and qualitative assessment (and consequently also perceptions and degree of systematicity when assessing) in the English as a Foreign Language (EFL) pronunciation class in tertiary education. Accordingly, the main objectives of this study have been to measure, rank and compare the harshness or leniency of the three different types of raters involved and their consistency/systematicity when in this role, as well as to measure the levels of coincidence and/or discrepancy when evaluating from different roles and depending on quantitative or qualitative considerations. The method used for the analysis involves the quantitative and qualitative analyses of the scores obtained in a triple-role assessment task carried out by 16 students and a teacher. The calculation of various statistical measures, the triangulation of the data obtained and the Many-Facet Rasch Measurement analysis of the results have completed the study and constitute the departure point for further larger studies. From the results obtained, it can be highlighted that quantitatively, the self and the teacher's perceptions seem to be the ones that are more distant or different, whereas the self and the peer's tend to be the most similar. In the same way, qualitative assessments seem to be more lenient, that is to say, slightly higher in mean score and have a lower coefficient of variation than quantitative ones in the three groups analysed. Consistency/systematicity is relatively high but it is still an aspect to be improved on the part of most raters.

Key words: Self-assessment, Peer-assessment, Teacher Assessment, Systematicity, Perceptions

INTRODUCTION

Pronunciation, as a key component of communicability, is a critical aspect in determining foreign language competence as well as in shaping others' perceptions of our own language competence. However, despite its determining role, it is probably one of the hardest aspects of orality to evaluate due to the many different variables which may not be perceived as the same or considered as equally important by raters and listeners in general. Such variables include accent, mother language interference, degree of deviation from what is considered native and/or "standard", and so forth. There are thus many perception-related issues that may considerably affect, either consciously or not, our perception of what is good pronunciation. The way we perceive a person's pronunciation and prefer it over another individual's way of pronouncing may even be determined by personal aspects (of which we may be more or less aware), such as sounding similar to somebody we love or like or being particularly interested in certain aspects such as tone, pitch, intonation, sound discrimination, etc. It would not be surprising then, for instance, that a pronunciation teacher does not perceive and evaluate pronunciation in the same way and with the same criteria as a student evaluates a peer or even him/herself. It is here where the issues of raters' perceptions and consistency arise, because since the origins of performance assessment human ratings have always been subject to various forms of error and bias, intrinsic to human nature. It is not unusual even for expert raters to come up with different ratings for the same performance, so that it seems that assessment largely depends upon which raters assign the rating (Eckes, 2015).

To shed some light on these aspects, to better understand if the same "assessed reality" is really differently qualitatively and quantitatively perceived by individuals and to determine if assessment is a unanimous or a controversial decision due to the human agents involved in it, the pilot study presented here is aimed at preliminarily describing, analysing, and comparing the way a tertiary class of pronunciation (with its teacher and 16 students) perceive and evaluate pronunciation from different perspectives and roles. This research has thus been designed and conducted as a "starting point" to provide an initial illustrative picture on assessment perceptions and systematicity which will be extended to a broader study –with a larger human sample and further research objectives– that is currently under development.

Published by Australian International Academic Centre PTY.LTD.

http://dx.doi.org/10.7575/aiac.ijalel.v.6n.7p.198

The main objectives of this study could thus be summarised as follows:

- To measure and analyse raters' levels of coincidence when evaluating from a quantitative or qualitative perspective and thus to determine whether one type of assessment tends to be harsher or more lenient than the other.
- To measure, analyse and draw conclusions on raters' levels of coincidence when evaluating from different roles (self (S) – peer (P) – teacher (T)); to compare raters' perceived level of pronunciation competence and thus their corresponding rating scores by analysing correlations between sets of marks.
- To be able to measure, rank and compare the harshness or leniency of the different types of raters and their consistency/systematicity when in this role;
- To see if systematicity in assessment and rubric interpretation can be expected and is possible when approached from different yet complementary roles.

Hence, this study aims to provide a comparative, descriptive, illustrative and reduced-but-representative picture of the way self, the peer and the teacher evaluate the same performance and thus on the way personal perceptions, rubric interpretation and consistency/systematicity as raters affect final scores. In this way, if (the aforementioned and other) further research in this area is promoted, possible causes for inconsistencies when evaluating pronunciation competence may be detected and corrected more easily, and systematicity and reliability in assessment through the use of shared and fair criteria and rubrics would be enhanced.

Theoretical Framework

Despite the challenges and difficulties implicit in any kind of assessment or assessment and despite the aforementioned "slippery" nature of pronunciation as an assessable entity, a number of interesting studies on pronunciation assessment and error-detection can be found (Neumeyer et al., 2000;Witt and Young, 2000;Cucchiarini et al., 2000; Franco et al., 2000; Moustroufas and Digalakis, 2007;Cincared et al., 2009; Striketal., 2009;Chen and Yang, 2012 and 2015, etc.). Most of these deal with the automatic assessment of pronunciation, even though there are also other studies encompassing, for instance, different but related topics such as the comparison/correlation of human and automatic scoring in pronunciation (Franco et al., 1997; Hacker et al., 2005).

Despite the difficulties of pronunciation scoring or assessment to provide an indication of the candidate's proficiency, this paper focuses on the assessment of pronunciation according to three different types of human raters' perceptions and criteria for comparison purposes. This author believes that agreement and coincidence in scores among different raters may not be as high as could be expected or advisable, but the results will have the last word.

Assessment roles and procedures

One of the biggest challenges in assessment is reaching a consensus or achieving consistency in results when different raters' scores are involved, apart from the well-known but blurry notion of fairness. The more the raters agree, the more reliable and consistent a mark seems, but this, although desirable and enriching, may not be easy to attain. This section sheds some light on the three rater roles (self, peer and teacher) participating in this study to better understand their criteria and validity as raters.

Klenowski (1995) defines self-assessment as "the assessment or judgement of 'the worth' of one's performance and the identification of one's strengths and weaknesses with a view to improving one's learning outcomes" (p. 146). The growth of students' ability to be realistic judges of their own performance together with the ability to monitor their own learning is one of the most important processes that can occur in undergraduate education (Boud and Lublin, 1983). Stefani (1994: 69) also agrees that "students have a realistic perception of their own abilities and can make rational judgements on the achievements of their peers". Likewise, Magin and Churches (1989) somehow agree with this idea by stating that developing future professionals' ability to assess and evaluate their own work at present but also in ways which can be applied to their future profession should be an important concern. It is thus agreed that self-assessment plays a critical role in the learning process in that it contributes to develop critical reflective practices in students which can subsequently be applied to their own peers and constitute hands-on practice for their future career as professionals.

As previous literature on the topic seems to indicate, taking into consideration student/candidate's self-assessment and setting common grounds as regards the assessment criteria of a subject or skill can only benefit the main actors in the assessment process. Incorporating self-assessment into regular class practices may result in an increased level of empathy between teachers and students, as well as in an increased critical and discerning capacity on both parts. Additionally, regular raters –normally teachers– should be able to fine-tune and improve their assessment schemes and criteria according to students' perceptions on them and, as Ross (2006) concedes, differences between self- and teacher-assessment can lead to productive teacher-student conversations about students' learning needs.

As regards peers' assessment, peer-assessment – understood as "an arrangement in which individuals consider the amount, level, value, worth and quality of success of the products or outcomes of learning of peers of similar status" (Topping, 1998: 250) – relies heavily on the judgement and objectivity of the students involved, which makes it necessary to implement it thoughtfully and cautiously (Frankland, 2007). Recent research on peer-assessment has been focused on its validity, which, in this author's view, is determined (as any other kind of assessment) by objectivity, coherence, consistency, systematicity, confidence, comfort and critical capacity.

It may be discussed whether, as Stefani (1994) states, some students may misuse their own power as raters by under-marking peers with the final aim of avoiding competition or simply giving themselves an advantage. What seems real is that probably the same parameters do not apply when teachers evaluate students and when students evaluate themselves or other peers, which should not imply either that one assessment is more valid or better than the other – they are simply different and sustained on different standards or motivations that very often raters themselves are not even aware of. As human beings we tend to perceive things differently and so our criteria (or even our interpretation of such criteria when these are pre-determined, as in rubrics) will possibly be different when evaluating the same phenomenon from two different role perspectives. In this regard, an interesting aspect for this study is the fact that some researchers (Boud and Middleton, 2003; Cheng and Warren, 2005) signal significant differences in the rating given by the teacher and the peer. In fact, according to Cheng and Warren (2005), students frequently report a low level of both comfort and confidence in their ability to fairly and responsibly assess their peers' proficiency. In fact, despite its importance and benefits for the learning process, according to Smith et al. (2002), some students (even though just a minority) still remain resistant to the principles and process of peer marking, due mainly to a lack of confidence in the ability of their peers to award fair and unbiased marks. This is so because of the many concerns students normally have about the fairness of peer-assessment, since they consider that other peers/students do not have adequate experience to judge them. In the same way, some other students, in an exercise of excessive self-criticism, also question their own capacity as raters, normally because of their undergraduate condition.

Nonetheless, as Cheng and Warren (2005) also concede, peer assessment is becoming an important alternative assessment method and, in this author's opinion, should be considered especially interesting and fruitful when compared and interpreted together with other types of "complementary" assessment. In Frankland's view (2007), however, another problem arises when the validity and reliability of peer assessment is often compared and judged according to the grade given by the tutor. The tutor is, supposedly, the most equitable standard or reference point, especially for the student, even when, as Falchikov et al. (2000) state, there is also uncertainty about teacher reliability and validity. Normally, the lecturer's power and legitimacy to evaluate is beyond question and students accept the established rules according to which it is the teacher who finally determines whether a student's grade is sufficient to pass or not. Nonetheless, it is very common to see how students' views on their own performance when evaluated may not coincide with those of the teacher. It is then when students claim that their mark is unfair, that it should be higher or that the lecturer has only taken into account X but Y was also important, and so on. In fact, if two people objectively score the same activity in a very different way, then we will probably agree that there is a problem that needs to be reflected upon in order to solve it. Almost the entire education system considers lecturers as the "most qualified" to evaluate, even when, as humans, they do commit mistakes and can/should also be trained, if necessary, to improve their assessment skills to make them more systematic and fair. Additionally, the same should happen with students so that they better understand the need to consistently apply assessment criteria and procedures as well as the discrimination techniques that can be employed to assign a mark. Even though it does not seem very feasible nowadays in the stressful routine we live in, the combination of a triple-perspective (self, peer and teacher) in assessment would most probably provide more complete and fairer results by adding a wider perspective to the overall assessment of a candidate.

The issue in this paper, however, is not so much to determine/acknowledge (again) the indisputable importance of being able to evaluate one's own or a peer's performance as fairly and consistently as possible but to ascertain up to what point students' and teacher's perceptions are coincident or partly shared as regards assessment criteria, as a departure point for further studies. In this author's opinion, this can only be checked by establishing to what extent there is a correlation between sets of marks assigned by different people when evaluating the same candidate and the same assessment evidence (task), something which is clearly delimited by perception issues.

Perception and rubric interpretation: two key aspects for defining "good pronunciation"

Pronunciation is probably one of the hardest aspects to evaluate when assessing an English as a Foreign Language (EFL) student. The raters' own perception of what constitutes "good pronunciation" may vary greatly from one to another and can be conditioned by previously acquired linguistic experience and even unconscious beliefs. In fact, the issue of how previously acquired linguistic experience shapes perception has been a relevant and prolific area of research over the past decades. Consequently, several comprehensive theories of speech perception of different natures have been developed (Acoustic Landmarks and Distinctive Features (Stevens, 2002), Exemplar theory (Johnson, 1996), among many others). As Carey et al. state (2011), these theories offer alternative explanations for the complex psychoacoustic processes underlying perception, and they can be said to collectively legitimise the part linguistic experience, or familiarity, plays in shaping perception. In fact, these authors refer to the self-coined concept of *interlanguage phonology familiarity* to try to justify their belief that the examiner's impression of the examinee's performance can be positively or negatively influenced according to the examiner's amount and type of exposure to the candidate's accent. Raters' familiarity with the accent of the candidate is thus a key aspect consciously or unconsciously affecting raters' perceptions, assessment criteria and thus final scores.

It is assumed that any rater will try to judge a candidate's performance on fair, objective, systematic and coherent criteria. Nonetheless, according to Carey et al. (2011:205), "despite the examiner's intentions to judge the candidate purely on the wording of the assessment criteria descriptors, the examiner's type and degree of L2 exposure could compete with the objectivity of the rating". This is so because, as these authors go onto state, want it or not, objectivity in rating is an unattainable ideal originated in the fact that the criterion-based rating and the speech perception of the rater are not directly related.

The role of the native speaker (NS) model is another aspect of perception which has played a controversial role in the assessment of pronunciation. As Taylor (2006) concedes, it was often assumed in the past that all language proficiency tests judge L2 performance against a "native speaker" criterion. However, it is also true that nowadays most tests no longer make any reference to native speaker competence among their assessment criteria or rating scales. This is probably due to the difficulty mentioned by Davies (2003) of how to really define this native speaker model accurately: Which NS? Where from? How young/old? What level of education?

Together with these aspects, the focus on accuracy or "correctness" also seems to have evolved into a more communicative reality: nowadays, as Taylor (2006) states, language assessment seems to have moved away from the traditional "deficit" model, based on penalising how "far away" someone is from the accepted standard, to current assessment criteria and performance descriptors, more focused on what someone can do. This fact makes the steady shift in language teaching/learning approaches more evident, as they move away from a focus on knowledge and form towards a focus on function and communication. Therefore, one important issue addressed by researchers concerns the relationship between perception and production. Numerous studies suggest that many L2 production difficulties are rooted in perception. Evidence also indicates that appropriate perceptual training can lead to automatic improvement in production and, most probably, consequently, also enhance systematicity, consistency and fairness in assessment. Precisely in order to add reliability and validity to assessments and to unify assessment criteria as much as possible in order to enhance them, the use of scoring rubrics is common practice today. According to Jonsson and Svingby (2007), the reliable scoring of performance can be fostered by the use of rubrics, especially if they are analytic, topic-specific and complemented with rater training. As they go on to state, rubrics seem to have the potential to promote learning and/or improve instruction by making expectations and criteria explicit, which facilitates feedback and self-assessment. Nonetheless, the use of a rubric cannot avoid the fact that a given performance is perceived differently by two raters; even the interpretation of rubric items may be different. The "product" to be evaluated and the set of criteria established to evaluate it will be the same, but not the minds undertaking the task. One could argue that the more detailed a rubric is, the better and more reliable or systematic its resulting assessments will be, but this may also be counterproductive. It is not feasible or pedagogical to use extremely long or detailed rubrics that prevent raters from being equally and sufficiently attentive to both the candidate's performance and the completion of the many-itemed rubric. Even in the case of using milimetrically-depicted items in a rubric, the rater's interpretations will always be present somehow, since perception and the interpretation of reality are subjective features inherent to human nature, regardless of the amount of detail provided. All in all, the transparency of assessment criteria provided by rubrics when making criteria explicit can be seen as a great contributor to learning, greatly enhancing feedback production, enhancing fairness and systematicity, and setting common grounds for the assessment process.

METHOD

The design of any study constitutes the blueprint for collecting, measuring and analysing relevant and revealing data. This study seeks to describe, quantify and compare self, peer and teacher assessments to better understand their perceptions, criteria and systematicity as raters. The resulting data obtained originate thus in the triangulation of the data obtained from the triple assessment proposed and respond to both quantitative and qualitative analyses combined with a series of complementary statistical measurements and calculations. Therefore, the research design adopted, empirical and contrastive/comparative, provides not just a description but also a comparison of the sets of data obtained, drawing a comprehensive picture on relevant discrepancies and similitudes between raters.

The procedure adopted was the following one: sixteen first-year Spanish university students, who were supposed to have a B1 level according to the Common European Framework of Reference for Languages (CEFR) and were enrolled in the subject "Pronunciation and comprehension of the English Language" participated[1] in this pilot study, together with their lecturer. All the participants (teacher included) had Spanish or Catalan as their mother tongue except one half-native student whose father was Irish. The data used in this research were compiled in a 2-hour session of the aforementioned subject. To be more specific, the method used for the analysis involved the following stages:

Firstly, the author of this study designed five documents, which were created in order to generate and compile assessment data:

1. Two documents with general instructions and final results grids (Appendix I):
   - A "General instructions and final results sheet (student version)"
   - A "General instructions and final results sheet (teacher version)"
2. Two assessment rubrics (Appendix II):
   - A "Quantitative assessment rubric or numerical rubric"
   - A "Qualitative assessment rubric or band rubric"
3. Tasks to be recorded on the part of students (versions A, B and C) (Appendix III).

Once the documents had been elaborated and during one of their regular classes of the subject "Pronunciation and comprehension of the English Language", students were asked to use their computers to record (in the language lab where the class normally took place) the pronunciation task designed in the previous stage (Appendix III). Three different versions (A, B and C) of the task to be recorded were created in order to avoid interference or imitated performance from students sitting nearby. The three tasks were thus different as regards content but similar regarding level and task type and students were handed out the tasks to be recorded (alternating versions A, B and C) so that those sitting side-by-side did not have the same version. The activities were designed by the author and comprised a free-speech task, a text-reading task and a sentence-reading task in order to allow the key pronunciation-related aspects (included in the rubrics) to be effectively evaluated from the different perspectives offered by the different tasks proposed. After finishing with their recordings, students uploaded their tasks in the virtual/online platform of the subject. This recording and uploading routine was familiar to them, since it was frequently used during their regular lessons.

Once the recordings had finished, two copies of the numerical rubric and one copy of the band rubric were given to each student participating in the research. The numerical rubric was designed entirely by the author of this study taking into account the main descriptors involved in pronunciation competence. It offers a more item-detailed kind of assessment. The band rubric was even more detailed but was intended to provide a holistic assessment, since it involved rating all the aspects considered in an integrated way, that is, with a letter representing the integrated performance of the different aspects specified. Therefore, since this rubric provided a holistic but detailed qualitative assessment, the values of these letters needed to be perfectly described and, thus, be comprehensible for raters. This band rubric was adapted (basically by simplifying it slightly and changing its format) from the speaking band descriptors (public version) employed by the International English Language Testing System (found athttp://ielts-yasi.englishlab.net/DETAILED_BAND_SCORE_DESCRIPTORS.htm)

The rubrics were used as follows:

- One copy of the numerical rubric was used for students' self-assessment; the other copy was used for the assessment of a peer assigned at random by the author (Appendix II).
- One copy of the band rubric was used both for self-assessment and peer-assessment; the first column in the rubric was devoted to self-assessment and the second column was used for the assessment of a peer assigned at random by the author (Appendix II).

Students were then asked to carefully read the document entitled "General instructions and final results sheet (student version)" (Appendix I). They were told that it was essential to understand and carefully follow the instructions given. Students could also pose any questions or doubts they might have whenever necessary.

At the same time, the "General instructions and final results sheet (teacher version)" (Appendix I) was given to the teacher for him/her to read attentively. Sixteen copies of the numerical rubric and sixteen copies of the band rubric were also provided to him/her in order to evaluate all the candidates both qualitatively and quantitatively. The teacher was allowed some additional days to be able to undertake the double (qualitative and quantitative) assessment of the sixteen students participating.

The pairs of peers to be evaluated were established at random: from the virtual/online platform of the subject where students had uploaded their recordings (and while students were reading the instructions and rubrics they had to work with), the author randomly assigned one peer to be reviewed per candidate. They received an anonymous recorded task via mail to evaluate in the same class. The name of the recorded candidate was substituted by a code whose name correspondence was only known to the author of this study. Even though I am aware of the fact that students could recognise their partners because of their voice, the assessment was carried out in a completely anonymous way and, unless the rater him/herself revealed it, the evaluated candidate would never know who evaluated him/her.

Students then proceeded to listen attentively to and evaluate their own recording and their assigned peer recording. In this way, every candidate's performance in the recorded pronunciation activity was evaluated by means of the two rubrics. This assessment was undertaken from a triple perspective: the candidate evaluated his/her own performance, then he/she evaluated a peer's performance, and finally the teacher evaluated all the candidates' performance. In this way, every recorded sample and thus every candidate was self-evaluated, peer-evaluated and teacher-evaluated both quantitatively and qualitatively for multi-perspective comparison purposes.

Students filled in the assessment results in the corresponding grid of the document "General instructions and final results sheet (student version)" and the teacher did the same in the "General instructions and final results sheet (teacher version)". Once all the resulting assessment sheets had been gathered and all the data compiled in tables for each of the assessments carried out (self, peer and teacher), a series of statistical scores relevant to the objectives of our research were calculated for comparison purposes (see section 3.1). These measures used for the analysis of both qualitative and quantitative results were of two main kinds:

- Measures of central tendency: mean and median. These are statistical terms with a somewhat similar role in the understanding of the central tendency of a set of statistical scores. However, despite the popularity of the mean as a measure of a mid-point in a sample, it has the disadvantage of being affected by any single value being too high or too low compared to the rest of the sample. This is why the median is sometimes taken as a better measure of a mid-point and why we have also incorporated it into our analysis.
- Measures of dispersion or spread: standard deviation[2] and coefficient of variation[3]. These were calculated in order to determine and evaluate the dispersion of the sample, since measures of central tendency estimate "normal" values of a dataset but do not describe the spread of the data or its variation around a central value as do measures of dispersion. Therefore, a proper description of a set of data should include both of these characteristics.

The data obtained from the raters was put in order and meaningfully arranged. In order to obtain comparable results/scores, qualitative assessment needed to be numerically interpreted/"translated". With this aim, a band correspondence was created for the interpretation of the numerical qualitative assessment interpretation (see Table 1). In this band correspondence, a series of numerical mean scores were calculated and established for each band. These calculated scores thus constituted a single, pre-determined figure indicating the mean value of the score range (of 0.99 points) included in each band. Consequently, these "quantitatively/numerically translated" scores necessarily constitute more restricted values than purely quantitative ones, where raters could choose the exact numerical score they wanted to assign but, as explained throughout the study, this aspect has been taken into consideration when putting forward the main results and conclusions of the research. Once the scores had been gathered and made compatible, calculations were undertaken and results analysed.

Table 1. Band correspondence for numerical QL assessment interpretation/"translation"Band for QL assessment

	Score range comprehended in each band	Corresponding numerical mean score for each band (established for comparative purposes)
A (best)	9	9
B	8–8.99	8.495
C	7–7.99	7.495
D	6–6.99	6.495
E	5–5.99	5.495
F	4–4.99	4.495
G	3–3.99	3.495
H	2–2.99	2.495
I (worst)	1–1.99	1.495

Finally, in order to provide an overall description of our set of data, the results/scores obtained according to the rubrics were also converted into a Many-Facet Rasch Measurement program-compatible format. Despite not being the main focus of this study, the data obtained were also analysed with this software (see section 3.2) so that further results could be retrieved in order to attain the objectives posed and complete the study from a different perspective.

RESULTS AND DISCUSSION

Statistical Analysis Results

One of the main objectives set for the study was to measure and compare raters' (self, peer and teacher) coincidence levels when evaluating from a quantitative (QN) or qualitative (QL) perspective and thus to determine whether one type of assessment tends to be harsher or more lenient than the other, also depending on the kind of rater. As mentioned in the previous section, with this aim, both measures of central tendency (mean and median) and measures of dispersion or spread (standard deviation and coefficient of dispersion) have been considered in order to provide a complementary and complete view on the topic.

Table 2 shows the sixteen candidates participating in the pilot study ("Candidate" column) as well as the three raters (self, peer, teacher) evaluating their performance in each case ("Rater" column). In this case, although numerical codes (1 to 16) have been maintained, names have been made public to better show the rater-candidate combinations. The QN and QL scores assigned to each candidate by each rater are shown in the "QN Score" and "QL Score"

Table 2. Candidates, raters, their corresponding QN and QL scores, and total counts

Candidate	Rater	QN Score	QL Score
1. Elia	Elia	4.8	C-7.495
Elia	Sofia	6.4	E-5.495
Elia	Teacher	3	F-4.495
2. Beatriz	Beatriz	6.2	C-7.495
Beatriz	Ioana	7.2	C-7.495
Beatriz	Teacher	7.4	C-7.495
3. Ioulia	Ioulia	6.5	B-8.495
Ioulia	Daniela	7.4	C-7.495
Ioulia	Teacher	7.7	B-8.495
4. Patricia	Patricia	6.5	D-6.495
Patricia	Irene	7.8	B-8.495
Patricia	Teacher	7.6	B-8.495
5. Bart	Bart	6.5	C-7.495
Bart	Paloma	7.2	C-7.495
Bart	Teacher	8.4	B-8.495
6. Irene	Irene	6.1	D-6.495
Irene	Nieves	6	C-7.495
Irene	Teacher	4.8	E-5.495

(*Contd...*)

Table 2. (***Contined***)

Candidate	Rater	QN Score	QL Score
7. Jose Ramon	Jose Ramon	6.6	D-6.495
Jose Ramon	Evelina	5	E-5.495
Jose Ramon	Teacher	5.6	D-6.495
8. Nieves	Nieves	6.1	D-6.495
Nieves	Andrea	6.2	D-6.495
Nieves	Teacher	3.5	F-4.495
9. Daniela	Daniela	5	E-5.495
Daniela	Ioulia	5.5	B-8.495
Daniela	Teacher	4.5	D-6.495
10. Paula	Paula	6.3	C-7.495
Paula	Elia	5	D-6.495
Paula	Teacher	3.7	F-4.495
11. Sofía	Sofía	6.8	D-6.495
Sofía	Bart	6.2	D-6.495
Sofía	Teacher	5.6	D-6.495
12. Ioana	Ioana	6.4	C-7.495
Ioana	Patricia	5.1	D-6.495
Ioana	Teacher	4.6	E-5.495
13. Andrea	Andrea	6.3	C-7.495
Andrea	Paula	7.8	B-8.495
Andrea	Teacher	6.4	D-6.495
14. Evelina	Evelina	5	E-5.495
Evelina	Ashley	7.5	C-7.495
Evelina	Teacher	3.4	F-4.495
15. Ashley	Ashley	8.5	A-9
Ashley	Beatriz	7.5	B-8.495
Ashley	Teacher	9	A-9
16. Paloma	Paloma	7	D-6.495
Paloma	Jose Ramon	7.4	C-7.495
Paloma	Teacher	5.6	E-5.495
Total number of students: 16	Total number of assessments: 48	Total QN score: 296.6	Total QL score: 330.77 (A=2; B=8; C=13; D=14; E=7; F=4)

columns respectively. In this way, for instance, in the case of candidate 1, named Elia, the assessment was carried out in the following way: firstly, Elia evaluated herself; secondly, Elia was evaluated by a peer called Sofía; and, thirdly, Elia was evaluated by the teacher. In each of the rater-candidate pairs, both QN and QL scores were provided by the rater, although in the QL Score the candidate provided the letter value of his/her choice (according to the qualitative assessment rubric in Appendix II) and the author "translated" it numerically according to Table 1 in order to provide comparable numerical data.

In this sense, and according to the data contained in Table 2, a first, general, comparison between quantitative and qualitative mean assessments results in the following:

QN ev.: 296.6 total score/48 assessments = 6.18

QL ev.: 330.77 total score/48 assessments = 6.9

In general terms, these preliminary results suggest that qualitative assessments seem to be more lenient, that is to say, slightly higher in mean score, than quantitative ones. If more detail is provided, Table 3 shows initial, general mean results about the harshness or leniency of the three different types of assessment conducted (QL or QN) and the differences between raters' roles:

Table 3. General results regarding the type of assessment (QN or QL) and the type of rater

Candidate	Rater	Mean QN score	Mean QL score
1 to 16	Self	6.29	7.03
1 to 16	Peer	6.58	7.25
1 to 16	Teacher	5.68	6.4

If the mean-based results in Table 3 are interpreted and ordered from more lenient to harsher, we find that the softest kind of assessment (therefore the one with the highest mean score) seems to be the qualitative assessment on the part of peers (7.25), whereas the hardest one (the one with the lowest mean score) seems to be the quantitative assessment on the part of the teacher (5.68). If arranged comprehensively in descending order, from more lenient to harsher kinds of assessment, we find:

- Peer qualitative assessment (P QL): 7.25 → Most lenient assessment
- Self qualitative assessment (S QL): 7.03
- Peer quantitative assessment (P QN): 6.58
- Teacher qualitative assessment (T QL): 6.4
- Self quantitative assessment (S QN): 6.29
- Teacher quantitative assessment (T QN): 5.68 → Harshest assessment

From all the discrepancies between qualitative and quantitative marks, only candidate 16 in self-assessment and candidate 1 in peer-assessment obtain a higher quantitative than qualitative mark (7% of the total amount of discrepancies); in the remaining 93% of discrepancies, the qualitative mark was higher.

If the rest of the measures considered are incorporated to complete the picture, then we get the results compiled in Table 4 (below). This table gathers central tendency results (not just the mean – already analysed above– but also the median) as well as measures of dispersion results. Median results have been calculated to reinforce the conclusions offered by mean values due to the fact that the mean, especially in small samples, may be affected by any single value being too high or too low compared to the rest of the sample. In this case, median measure results corroborate that, if we compare the data of the same population measured or evaluated in the three different ways (self, peer and teacher), the quantitative assessment carried out by the teacher is the lowest (harshest) one (median = 5.6) if compared with self-assessment (median 6.35) and peer-assessment (median = 6.8), exactly

Table 4. Levels of coincidence between QN and QL assessments on the part of the different kinds of raters participating in the study. (N=No; Y=Yes)

Candidate	Self-assessment SELF-EV. (mark given by the candidate to him/ herself)		Do QN& QL fully coincide?	Peer-assessment PEER-EV. (mark given by the peer to the candidate)		Do QN& QL fully coincide?	Teacher's assessment TEACHER EV. (mark given by the teacher to the candidate)		Do QN& QL fully coincide?
	Number	**Band**		**Number**	**Band**		**Number**	**Band**	
Elia	4.8	C	N	6.4	E	N	3	F	N
Beatriz	6.2	C	N	7.2	C	Y	7.4	C	Y
Ioulia	6.5	B	N	7.4	C	Y	7.7	B	N
Patricia	6.5	D	Y	7.8	B	N	7.6	B	N
Bart	6.5	C	N	7.2	C	Y	8.4	B	Y
Irene	6.1	D	Y	6	C	N	4.8	E	N
José Ramón	6.6	D	Y	5	E	Y	5.6	D	N
Nieves	6.1	D	Y	6.2	D	Y	3.5	F	N
Daniela	5	E	Y	5.5	B	N	4.5	D	N
Paula	6.3	C	N	5	D	N	3.7	F	N
Sofía	6.8	D	Y	6.2	D	Y	5.6	D	N
Ioana	6.4	C	N	5.1	D	N	4.6	E	N
Andrea	6.3	C	N	7.88	B	N	6.4	D	Y
Evelina	5	E	Y	7.55	C	Y	3.4	F	N
Ashley	8.55	A	N	7.5	B	N	9	A	Y
Paloma	7	D	N	7.4	C	Y	5.6	E	Y
Median	6.35	C-D		6.8	C		5.6	D	
MEAN (also in Table 3)	6.29	7.03		6.58	7.25		5.68	6.40	
Standard Deviation	0.88	0.96		1.04	1.00		1.90	1.59	
Coefficient of Variatoin	14%	14%		16%	14%		33%	25%	
Coincidences between QN and QL assessments			7/16			8/16			5/16

the same as what occurs with qualitative assessment (and also coinciding with mean results). If further interpretation of results is provided, it may be observed that peer-assessment has the highest median value compared to the other two values. Moreover, none of the values is less than five points, which may lead us to think that the assessment is not totally objective and that students probably do not want to assume the responsibility of a partner failing. Moreover, in all three groups qualitative marks are higher than quantitative ones, as Table 3, in general[4], also reflected through mean values.

For dispersion to be considered, the "standard deviation" and "coefficient of variation" lines in Table 4 need to be taken into account. These lines contain all the information necessary to understand the dispersion in our sample more comprehensively and graphically, since the central tendency measures analysed so far do not describe data variation around a central value or the spread of the data as do measures of dispersion. Therefore, a proper description of a set of data should include both of these characteristics. Hence, Table 4 shows that the teacher's coefficient of variation is the highest one both in the quantitative (33%) and qualitative (25%) measures. Therefore, although teacher assessment is the lowest on measuring both qualitatively and quantitatively, the fact that it possesses the highest coefficient of variation means that there is a greater dispersion, that is, the teacher's assessments are the ones showing the highest amount of variability relative to the mean. This increase in dispersion is due to the fact that there are marks which are closer to 9 and 3, something which does not happen in the other two groups. Presumably, the assumed greater linguistic expertise of the teacher makes him/her capable of distinguishing students' lack of precision or difficulties more accurately and his/her teacher role makes him/her more confident when scoring and even when using more "extreme" values (closer to 9 and to 3). Not fully coinciding with what Langan et al. (2008) found – that in peer-assessment students over-marked colleagues probably because of the close-knit community created –, in this study, however, pupils seem to be more cautious and prefer to give both their peers and themselves more moderate scores, probably also influenced by their assumed lack of expertise when doing this task and trying to reach a balance that avoids possible unfair assessments. Thus, quantitatively, self-assessment is the one showing the lowest dispersion value (14%), whereas peer-assessment has a dispersion of 16%. The same occurs if we consider qualitative assessment, since the dispersion of self-assessment and peer-assessment is 14%, which is considerably inferior to the 25% obtained in the teacher-assessment group. This aforementioned self-assessment, with a median of 6.35 points and the lowest dispersion value, should make us notice that all the self-assessment marks (except one of them because of two tenths) are equal to or more than 5 points, which can indicate that students do not see defects objectively, but they are not capable of highlighting virtues either.

If the coefficients of variation of qualitative and quantitative assessments are compared, we observe that qualitative marks have a lower coefficient of variation than quantitative ones in the three groups. At this point, it should be noted that the dispersion of quantitative marks will always be higher than that of qualitative ones for a simple reason: from 1 to 10 there are infinite values whereas from A to I there are only 10 possible values. The same will happen if we use the mean of the range of values calculated in Table 1. *A priori*, the expected dispersion of the sixteen marks on comparing qualitative and quantitative ones will always be lower in qualitative ones. If we observe the standard deviation values of the three measures, we can again see how qualitative and quantitative dispersion is very similar except in the case of teacher assessment. It is also rather surprising that in self-assessment the coefficient of variation is lower in quantitative than in qualitative data, where 11 out of 16 values are around a mean mark of 6. This is why, in order to compare the measures among them, it is more correct to use the coefficient of variation (Mean/standard deviation) in order to value the population as a whole. If the three measures measure the same (the level of pronunciation in English), the results indicate that self-assessment and peer assessment are more similar to each other than teacher assessment, which is different, with a lower mean and median and more dispersion. These findings are not fully coincident, however, with those of authors such as Li et al. (2006), according to which, peer ratings tend to show a moderately high level of agreement with teacher ratings. All in all, as Langman et al (2008: 187) contend, it seems out of question that "understanding the processes of self and peer assessment requires an appreciation of students' perceptions of themselves and others […]. Such understanding should enhance future implementations".

Another key aspect to analyse (and closely related to the spread or dispersion of data) is consistency and systematicity on the part of raters when evaluating quantitatively and qualitatively, that is to say, whether (or up to what point) they quantitatively and qualitatively rate candidates within the same score range (as established in Table 1) without knowing the numerical correspondence between qualitative-quantitative ratings in advance. The author is aware of the fact that the rubrics themselves may strongly affect the final results in this respect. It is a fact that the quantitative band is more visual and straight-to-the-point than the qualitative one, which is considerably longer and more detailed –harder to manage, we could say. It is also true that the researcher cannot individually control every single rater and guarantee that they carefully read, understand, consider and evaluate all the items in the qualitative rubric. The likely unbalanced importance assigned by the rater to each aspect in the holistic assessment of the band rubric is also something that cannot be controlled by the researcher. Nonetheless, despite these small contingencies that are always present in any kind of research, instructions were clearly explained, enough time was given to read and become familiar with the rubrics and, more importantly, the numerical and band rubrics were designed with the same variables (intonation, accent, understandability, sentence stress, etc.). As a result, the comparability of the final scores is real, as are the correspondences established in Table 1, and can reveal important data about the rater's consistency. With this aim in mind, firstly, a series of comparisons of single raters' QN and QL assessments were established in order to be able to respond to these initial questions:

- Do the QN and QL scores of the candidate coincide in self-assessment? That is to say, are the quantitative and qualitative marks assigned by the candidate to him/herself the same?
- Do the QN and QL scores of the candidate coincide in peer-assessment? That is to say, is the qualitative mark given by the peer to the candidate the same as the one given quantitatively?
- Do the QN and QL scores of the candidate coincide in teacher-assessment? That is to say, is the mark given qualitatively by the teacher to the candidate the same as the one given quantitatively?

Table 4 also preliminarily shows for each of the rater categories established (self, peer and teacher), and for each of the participating candidates, whether the quantitative scores assigned by raters show a logical correspondence or coincidence[5] with the qualitative marks (letters) also assigned by them. Results in Table 4 indicate that quantitative-qualitative "full-coincidences" (according to Table 1) in self, peer and teacher assessments are less frequent than expected, the lowest degree of coincidence corresponding to the teacher's ratings (just 31.25%, (5/16)), whereas in self-assessment the rate of coincidence is 43.75% (7/16) and 50% (8/16) in peer assessment. Despite not constituting a highly significant percentage/amount, peer assessment thus seems to be the most consistent one in terms of "full" correspondence/coincidence of quantitative and qualitative assessment, also probably influenced by students' "milder" character as raters.

In order to provide further detail and comparison on QN and QL consistency among raters and to do so in a graphical manner, Table 5 numerically calculates and illustrates the exact degree of correspondence or coincidence between QN and QL assessment in the following rating combinations, establishing levels of coincidence for the following specific assumptions:

S-P (self-peer):

- What coincidence level do we find when comparing self-QN-assessment and peer-QN-assessment? (Column (1), Table 5)
- What coincidence level do we find when comparing self-QL-assessment and peer-QL-assessmnt? (Column (5), Table 5)

S-T (self-teacher)

- What coincidence level do we find when comparing self-QN-assessment and teacher-QN-assessment? (Column (2), Table 5)
- What coincidence level do we find when comparing self-QL-assessment and teacher-QL-assessment? (Column (6), Table 5)

P-T (peer-teacher)

- What coincidence level do we find when comparing peer-QN-assessment and teacher-QN-assessment? (Column (3), Table 5)
- What coincidence level do we find when comparing peer-QL-assessment and teacher-QL-assessment?

Table 5. Correspondence or coincidence scores in the QN and QL assessments carried out by self (S), peer (P) and teacher (T)

	Numerical rubric score or QN assessment **Coincidence of +/–0.5 points (✓✓) between:** **Coincidence of +/–0.99 points (✓) between:** **No coincidence (X) between:**				**Band rubric score or QL assessment** **Coincidence of letter (✓✓) between:** **Coincidence of +/–1 letter (✓) between:** **No coincidence (X) between:**			
Candidate	**S-P (1)**	**S-T (2)**	**P-T (3)**	**S-P-T (4)**	**S-P (5)**	**S-T (6)**	**P-T (7)**	**S-P-T (8)**
1.	X	X	X	X	X	X	✓	X
2.	✓	X	✓✓	X	✓✓	✓✓	✓✓	✓✓
3.	✓	X	✓✓	X	✓	✓✓	✓	✓
4.	X	X	✓✓	X	X	X	✓✓	X
5.	✓	X	X	X	✓✓	✓	✓	✓
6.	✓✓	X	X	X	✓	✓	✓✓	✓
7.	X	✓	✓	X	✓	✓	✓✓	✓
8.	✓✓	X	X	X	✓✓	X	X	X
9.	✓✓	X	✓	X	X	✓	X	X
10.	X	X	X	X	X	X	X	X
11.	✓✓	X	✓	X	✓✓	✓✓	✓✓	✓✓
12.	X	X	✓✓	X	✓	X	✓	X
13.	✓	✓✓	✓	X	✓	✓	X	X
14.	X	X	X	X	X	✓	X	X
15.	✓	✓✓	X	X	✓	✓✓	✓	✓
16.	✓✓	X	X	X	✓	✓	X	X
Coincidences	5/16 ✓ 5/16 ✓✓	1/16 ✓ 2/16 ✓✓	4/16 ✓ 4/16 ✓✓	---	7/16 ✓ 4/16 ✓✓	7/16 ✓ 4/16 ✓✓	5/16 ✓ 5/16 ✓✓	5/16 ✓ 2/16✓✓

(Column (7), Table 5)

S-P-T (self, peer-teacher)

- What coincidence level do we find when comparing self-QN-assessment, peer-QN-assessment and teacher-QN-assessment? (Column (4), Table 5)
- What coincidence level do we find when comparing self-QL-assessment, peer-QL-assessment and teacher-QL-assessment? (Column (8), Table 5)

Results in Table 5 show that "full" score coincidences of +/- 0.5 points (represented as ✓✓ in the table) in numerical rubric scores (QN assessment) were the following:

- 31.25% (5/16) when comparing self and peer assessment scores.
- 12.5% (2/16) when comparing self and teacher assessment scores.
- 25% (4/16) when comparing peer and teacher assessment scores.
- None when comparing the three raters' role scores.

Results show that coincidences of +/- 0.99 points (represented as ✓ in the table) in numerical rubric scores (QN assessment) were the ones shown below. In the same way, this 0.99 points is the score range established for QN and QL correspondences in Table 1, so the author considers that two scores placed within this range or with a difference of up to 0.99 points can also be considered a coincidence to be taken into account.

- 31.25% (5/16) when comparing self and peer assessment scores.
- 6.25% (1/16) when comparing self and teacher assessment scores.
- 25% (4/16) when comparing peer and teacher assessment scores.
- None when comparing the three raters' role scores.

Results show that "full" coincidences of letter (represented as ✓✓ in the table) in band rubric scores (QL assessment) are the following:

- 25% (4/16) when comparing self and peer assessment scores.
- 25% (4/16) when comparing self and teacher assessment scores.
- 31.25% (5/16) when comparing peer and teacher assessment scores.
- 12.5% (2/16) when comparing the three raters' role scores.

Results show that "medium" coincidences of +/- 1 letter (represented as ✓in the table) in band rubric scores (QL assessment) are the following:

- 43.75% (7/16) when comparing self and peer assessment scores.
- 43.75% (7/16) when comparing self and teacher assessment scores.
- 31.25% (5/16) when comparing peer and teacher assessment scores.
- 31.25% (5/16) when comparing the three raters' role scores.

Qualitatively, we can see that coincidence levels are, in general, higher than in QN assessments but "full" correspondence or coincidence levels (✓✓) do not reach 50% in any of the instances analysed. However, if coincidence or consistency is approached more broadly, that is, considering both coincidence levels,(✓✓) and (✓),then there is a significant increase in coincidence measures, especially qualitatively speaking.

Coincidence levels of up to +/- 0.99 points, (✓✓) or (✓), in numerical rubric scores (QN assessment):

S-P: 10/16 → 62.5%
S-T: 3/16 → 18.7%
P-T: 8/16 → 50%
S-P-T: 0/16 → 0%

Coincidence levels of up to +/- 1 letter, (✓✓) or (✓), in band rubric scores (QL assessment):

S-P: 11/16 → 68.7%
S-T: 11/16 → 68.7%
P-T: 10/16 → 62.5%
S-P-T: 7/16 → 43.7%

Quantitatively, we can see that the self and the teacher's perceptions seem to be the ones that are more distant or different, whereas the self and the peers' ones tend to be the most similar ones. This is probably due to the fact that we are talking about the same profile of person: the candidate (self) is a student and the peer-rater is also another student from the same class and, presumably, with many similar features. Qualitatively, coincidence levels are, rather surprisingly, also much higher between self and teacher's views, which can indicate that the greater detail provided in the qualitative rubric makes it clearer and thus the criteria are more easily interpreted and agreed upon. In fact, as Miller (2003) states, the number and nature of criteria employed in assessment have a direct influence in the marks generated. In the same way, highly discordant levels at this point may also indicate a need to clarify or unify pronunciation assessment criteria and their interpretation in rubrics.

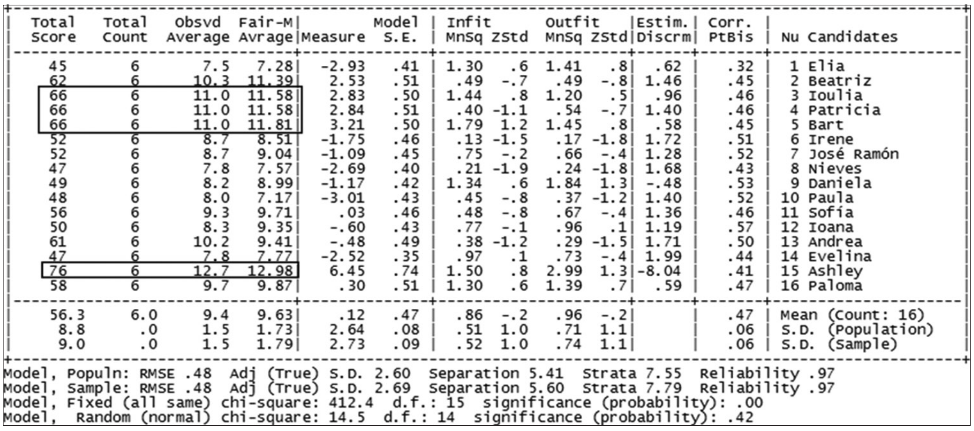
Finally, to complete and complement the study and provide a more comprehensive approach to the topic, correlations between sets of marks and raters' roles have been analysed with the software *Many-Facet Rasch Measurement* and the main results are presented in the next section.

Many-Facet Rasch Analysis Results

The Many-Facet Rasch Measurement (MFRM) is a psychometric approach which enables users to establish a coherent framework for drawing reliable, valid and fair inferences from rater-mediated assessments, thus answering the problem of fallible human ratings (Eckes, 2015). Even though it is not the central focus of this study, the reduced Facet Rasch analysis shown here is intended as a means to fine-tunepreviously provided data in a more individualised way so that all the research questions posed can be answered in a meaningful and comprehensive manner,and some additional but complementary data may be meaningfully incorporated.

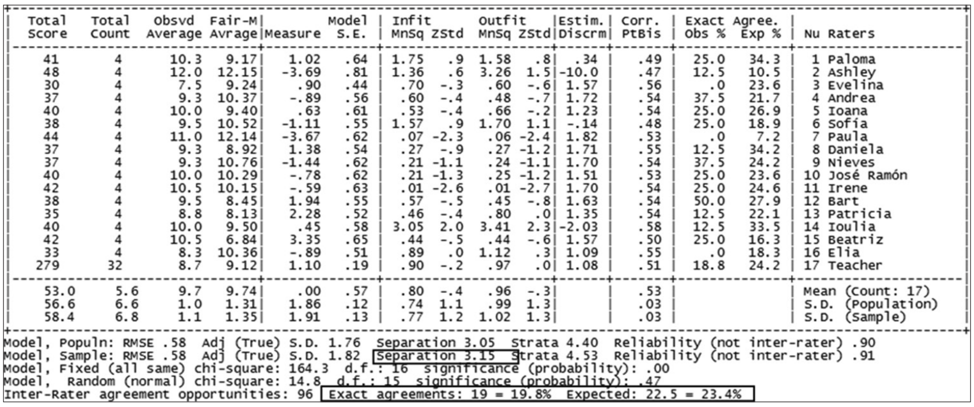
If fairness issues are addressed, Table 6 includes the Facet Rasch-generated ability measures for all candidates. We can highlight the fact that top-rated candidates –with a Total score of 76 (1 candidate) and 66 (3 candidates) – were under-evaluated by themselves, by their peer-rater and by the teacher, since they seemed

Table 6. Candidates' ability measures according to MFRM.



```
+-----------------------------------------------------------------------------------------------------------+
| Total   Total   Obsvd  Fair-M|        Model | Infit       Outfit    |Estim.| Corr. |                      |
| Score   Count  Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtBis | Nu Candidates        |
|------------------------------+--------------+---------------------+------+-------+----------------------|
|   45      6       7.5   7.28|  -2.93   .41 | 1.30   .6  1.41   .8|  .62 |   .32 |  1 Elia              |
|   62      6      10.3  11.39|   2.53   .51 |  .49  -.7   .49  -.8| 1.46 |   .45 |  2 Beatriz           |
|   66      6      11.0  11.58|   2.83   .50 | 1.44   .8  1.20   .5|  .96 |   .46 |  3 Ioulia            |
|   66      6      11.0  11.58|   2.84   .51 |  .40 -1.1   .54  -.7| 1.40 |   .46 |  4 Patricia          |
|   66      6      11.0  11.81|   3.21   .50 | 1.79  1.2  1.45   .8|  .58 |   .45 |  5 Bart              |
|   52      6       8.7   8.51|  -1.75   .46 |  .13 -1.5   .17 -1.8| 1.72 |   .51 |  6 Irene             |
|   52      6       8.7   9.04|  -1.09   .45 |  .75  -.2   .66  -.4| 1.28 |   .52 |  7 José Ramón        |
|   47      6       7.8   7.57|  -2.69   .40 |  .21 -1.9   .24 -1.8| 1.68 |   .43 |  8 Nieves            |
|   49      6       8.2   8.99|  -1.17   .42 | 1.34   .6  1.84  1.3| -.48 |   .53 |  9 Daniela           |
|   48      6       8.0   7.17|  -3.01   .43 |  .45  -.8   .37 -1.2| 1.40 |   .52 | 10 Paula             |
|   56      6       9.3   9.71|    .03   .46 |  .48  -.8   .67  -.4| 1.36 |   .46 | 11 Sofía             |
|   50      6       8.3   9.35|   -.60   .43 |  .77  -.1   .96   .1| 1.19 |   .57 | 12 Ioana             |
|   61      6      10.2   9.41|   -.48   .49 |  .38 -1.2   .29 -1.5| 1.71 |   .50 | 13 Andrea            |
|   47      6       7.8   7.77|  -2.52   .35 |  .97   .1   .73  -.4| 1.99 |   .44 | 14 Evelina           |
|   76      6      12.7  12.98|   6.45   .74 | 1.50   .8  2.99  1.3|-8.04 |   .41 | 15 Ashley            |
|   58      6       9.7   9.87|    .30   .51 | 1.30   .6  1.39   .7|  .59 |   .47 | 16 Paloma            |
|------------------------------+--------------+---------------------+------+-------+----------------------|
|   56.3    6.0     9.4   9.63|    .12   .47 |  .86  -.2   .96  -.2|      |   .47 | Mean (Count: 16)     |
|    8.8     .0     1.5   1.73|   2.64   .08 |  .51  1.0   .71  1.1|      |   .06 | S.D. (Population)    |
|    9.0     .0     1.5   1.79|   2.73   .09 |  .52  1.0   .74  1.1|      |   .06 | S.D. (Sample)        |
+-----------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .48  Adj (True) S.D. 2.60  Separation 5.41  Strata 7.55  Reliability .97
Model, Sample: RMSE .48  Adj (True) S.D. 2.69  Separation 5.60  Strata 7.79  Reliability .97
Model, Fixed (all same) chi-square: 412.4  d.f.: 15  significance (probability): .00
Model,  Random (normal) chi-square: 14.5  d.f.: 14  significance (probability): .42
```

Table 7. Rater performance and the severity measures for raters.



```
+-----------------------------------------------------------------------------------------------------------------+
| Total  Total  Obsvd  Fair-M|        Model | Infit      Outfit    |Estim.| Corr. | Exact Agree. |                   |
| Score  Count Average Avrage|Measure  S.E. | MnSq ZStd  MnSq ZStd|Discrm| PtBis | Obs %  Exp % | Nu Raters         |
|----------------------------+--------------+--------------------+------+-------+--------------+-------------------|
|   41     4     10.3   9.17|   1.02   .64 | 1.75   .9  1.58   .8|  .34 |   .49 |  25.0   34.3 |  1 Paloma         |
|   48     4     12.0  12.15|  -3.69   .81 | 1.36   .6  3.26  1.5|-10.0 |   .47 |  12.5   10.5 |  2 Ashley         |
|   30     4      7.5   9.24|    .90   .44 |  .70  -.3   .60  -.6| 1.57 |   .56 |    .0   23.6 |  3 Evelina        |
|   37     4      9.3  10.37|   -.89   .56 |  .60  -.4   .48  -.7| 1.72 |   .54 |  37.5   21.7 |  4 Andrea         |
|   40     4     10.0   9.40|    .63   .61 |  .53  -.4   .66  -.2| 1.23 |   .54 |  25.0   26.9 |  5 Ioana          |
|   38     4      9.5  10.52|  -1.11   .55 | 1.57   .9  1.70  1.1| -.14 |   .48 |  25.0   18.9 |  6 Sofía          |
|   44     4     11.0  12.14|  -3.67   .62 |  .07 -2.3   .06 -2.4| 1.82 |   .53 |    .0    7.2 |  7 Paula          |
|   37     4      9.3   8.92|   1.38   .54 |  .27  -.9   .27 -1.2| 1.71 |   .55 |  12.5   34.2 |  8 Daniela        |
|   37     4      9.3  10.76|  -1.44   .62 |  .21 -1.1   .24 -1.1| 1.70 |   .54 |  37.5   24.2 |  9 Nieves         |
|   40     4     10.0  10.29|   -.78   .62 |  .21 -1.3   .25 -1.2| 1.51 |   .53 |  25.0   23.6 | 10 José Ramón     |
|   42     4     10.5  10.15|   -.59   .63 |  .01 -2.6   .01 -2.7| 1.70 |   .54 |  25.0   24.6 | 11 Irene          |
|   38     4      9.5   8.45|   1.94   .55 |  .57  -.5   .45  -.8| 1.63 |   .54 |  50.0   27.9 | 12 Bart           |
|   35     4      8.8   8.13|   2.28   .52 |  .46  -.4   .80   .0| 1.35 |   .54 |  12.5   22.1 | 13 Patricia       |
|   40     4     10.0   9.50|    .45   .58 | 3.05  2.0  3.41  2.3|-2.03 |   .58 |  12.5   33.5 | 14 Ioulia         |
|   42     4     10.5   6.84|   3.35   .65 |  .44  -.5   .44  -.6| 1.57 |   .50 |  25.0   16.3 | 15 Beatriz        |
|   33     4      8.3  10.36|   -.89   .51 |  .89   .0  1.12   .3| 1.09 |   .55 |    .0   18.3 | 16 Elia           |
|  279    32      8.7   9.12|   1.10   .19 |  .90  -.2   .97   .0| 1.08 |   .51 |  18.8   24.2 | 17 Teacher        |
|----------------------------+--------------+--------------------+------+-------+--------------+-------------------|
|   53.0   5.6    9.7   9.74|    .00   .57 |  .80  -.4   .96  -.3|      |   .53 |              | Mean (Count: 17)  |
|   56.6   6.6    1.0   1.31|   1.86   .12 |  .74  1.1   .99  1.3|      |   .03 |              | S.D. (Population) |
|   58.4   6.8    1.1   1.35|   1.91   .13 |  .77  1.2  1.02  1.3|      |   .03 |              | S.D. (Sample)     |
+-----------------------------------------------------------------------------------------------------------------+
Model, Populn: RMSE .58  Adj (True) S.D. 1.76  Separation 3.05  Strata 4.40  Reliability (not inter-rater) .90
Model, Sample: RMSE .58  Adj (True) S.D. 1.82  Separation 3.15  Strata 4.53  Reliability (not inter-rater) .91
Model, Fixed (all same) chi-square: 164.3  d.f.: 16  significance (probability): .00
Model,  Random (normal) chi-square: 14.8  d.f.: 15  significance (probability): .47
Inter-Rater agreement opportunities: 96  Exact agreements: 19 = 19.8%  Expected: 22.5 = 23.4%
```

to deserve higher marks according to the Fair-M Average score generated by the program. Moreover, according to the results obtained, there are at least 5.60 statistically distinct performance levels in the sample (this fact having a high reliability of 0.97), which further sustains the distribution analysis of section 3.1.

The Model S.E. score shows the standard error, that is, how big or small the error is depending on the extent to which Observed Average and Fair-M Average coincide. In this particular instance we can see that the error is significant in certain cases, the most significant being the 0.74 standard error of the top-ranked candidate.

According to the InfitMnsq column, most candidates are being consistent, since their resulting scores are within the 0.5 to 1.5 range. Only in the cases of Beatriz (0.49), Patricia (0.40), Irene (0.13), Nieves (0.21), Paula (0.45), Sofía (0.48) and Andrea (0.38)is this consistency more dubious.

As may be observed in Table 6, since exact agreements are close to expected agreements, our raters could be described more as independent raters than as "scoring machines". At the same time, the high reliability value confirms that the rater measures are different. In the same way, high InfitMnsq values suggest rater inconsistency, that is, variation or noise in ratings, but this only happens with three candidates: Paloma (1.75), Sofía (1.57) and Ioulia (3.05).

Table 7 looks at rater performance and includes the severity measures for raters. From the results we may see that the most lenient rater was Ashley (who, curiously enough, is the only native speaker among the candidates and also the top-rated candidate) with a score in the Measure column of -3.69; the harshest rater was Beatriz (3.35). According to the program, the raters participating in the study can be divided into at least 3.15 statistically different severity levels.

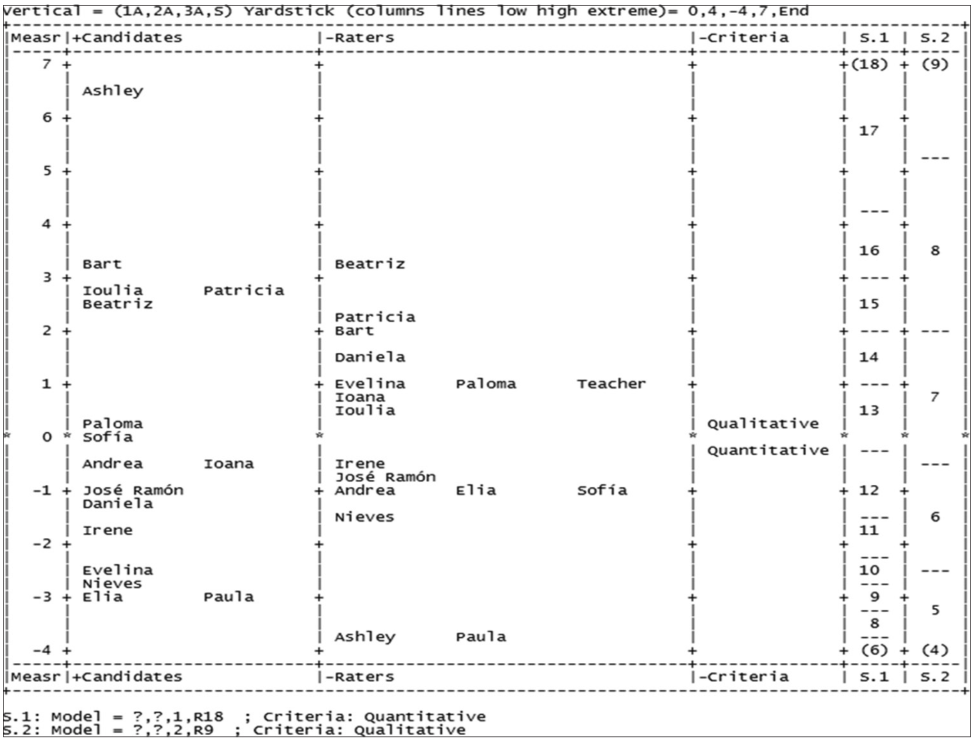
Finally, Table 8 graphically summarises the strongest (at the top) and the weakest (at the bottom) candidates and the harshest (at the top) and most lenient raters (at the bottom).

Results in this respect also show that raters are statistically different in terms of harshness and leniency but relatively self-consistent. In the same way, the two criteria employed (quantitative and qualitative) are relatively close in measure and not statistically different. Therefore, coincidences are higher than expected and no drastic discrepancies are observed as regards assessment criteria in the different roles analysed.

CONCLUSIONS

It is important that students are evaluated according to understandable and shared criteria which they can themselves reproduce and agree with. Developing student's ability to evaluate and enhancing the importance of correctly interpreting the criteria contained in rubrics or simply their own criteria and consistency as raters in order to fairly and

Table 8. Candidates and raters' harshness and leniency values



```
Vertical = (1A,2A,3A,S) Yardstick (columns lines low high extreme)= 0,4,-4,7,End
+------------------------------------------------------------------------------------------------+
|Measr|+Candidates              |-Raters                              |-Criteria     | S.1 | S.2 |
|-----+-------------------------+-------------------------------------+--------------+-----+-----|
|   7 +                         +                                     +              +(18) + (9) |
|     |                         |                                     |              |     |     |
|     | Ashley                  |                                     |              |     |     |
|     |                         |                                     |              |     |     |
|   6 +                         +                                     +              +     +     |
|     |                         |                                     |              | 17  |     |
|     |                         |                                     |              |     |     |
|     |                         |                                     |              |     | --- |
|   5 +                         +                                     +              +     +     |
|     |                         |                                     |              |     |     |
|     |                         |                                     |              |     |     |
|     |                         |                                     |              | --- |     |
|   4 +                         +                                     +              +     +     |
|     |                         |                                     |              |     |     |
|     |                         |                                     |              | 16  |  8  |
|     | Bart                    | Beatriz                             |              |     |     |
|   3 +                         +                                     +              + --- +     |
|     | Ioulia      Patricia    |                                     |              |     |     |
|     | Beatriz                 |                                     |              | 15  |     |
|     |                         | Patricia                            |              |     |     |
|   2 +                         + Bart                                +              + --- + --- |
|     |                         |                                     |              |     |     |
|     |                         | Daniela                             |              | 14  |     |
|     |                         |                                     |              |     |     |
|   1 +                         + Evelina     Paloma      Teacher     +              + --- +     |
|     |                         | Ioana                               |              |     |  7  |
|     |                         | Ioulia                              |              | 13  |     |
|     | Paloma                  |                                     | Qualitative  |     |     |
*   0 * Sofía                   *                                     *              *     *     *
|     |                         |                                     | Quantitative | --- |     |
|     | Andrea      Ioana       | Irene                               |              |     | --- |
|     |                         | José Ramón                          |              |     |     |
|  -1 + José Ramón              + Andrea      Elia        Sofía       +              + 12  +     |
|     | Daniela                 |                                     |              |     |     |
|     |                         | Nieves                              |              | --- |  6  |
|     | Irene                   |                                     |              | 11  |     |
|  -2 +                         +                                     +              +     +     |
|     |                         |                                     |              | --- |     |
|     | Evelina                 |                                     |              | 10  | --- |
|     | Nieves                  |                                     |              | --- |     |
|  -3 + Elia        Paula       +                                     +              +  9  +     |
|     |                         |                                     |              | --- |  5  |
|     |                         |                                     |              |  8  |     |
|     |                         | Ashley      Paula                   |              | --- |     |
|  -4 +                         +                                     +              + (6) + (4) |
|-----+-------------------------+-------------------------------------+--------------+-----+-----|
|Measr|+Candidates              |-Raters                              |-Criteria     | S.1 | S.2 |
+------------------------------------------------------------------------------------------------+

S.1: Model = ?,?,1,R18  ; Criteria: Quantitative
S.2: Model = ?,?,2,R9   ; Criteria: Qualitative
```

systematically evaluate their own and their peers' work is a crucial aspect when training future professionals and when trying to guarantee fairness in assessment processes.

Triple-role assessment can be a suitable assessment method combining the benefits of a triple-perspective assessment with the reliability of shared, comparable criteria. Nonetheless, apart from the time restrictions making this kind of assessment rather unfeasible to be carried out on a regular basis, it is common that students feel uncomfortable and even unqualified to assess their own and especially their peers' language proficiency. However, in fact, they are a very powerful "tool" to carry out assessments grounded on an ample combination of perspectives that make the assessment process fully operational, integrative and fair. Results show that there are differences (although maybe not as significant as one might think) when pronunciation proficiency is scored by the self, the peer or the tutor. In this sense, "assessment behaviours" tend to follow a general common path, especially among student-raters, the teacher being the one showing more discrepancies with respect to the former. Reaching that middle point of equilibrium in which students and teachers understand, share and perceive each other's criteria in the same way when scoring is the key to achieve the long-awaited fairness and systematicity in assessment. Whatever the case, any kind of assessment or assessment implies an amount of responsibility, coherence, systematicity and knowledge, which the author firmly believes can and should be trained for the sake of both teachers and students.

AKNOWLEGMENT

This research is part of the Education and Innovation Research Project: Proyecto de Innovación Educativa Universitat Jaume I 2779/13 Parámetros de aproximación a la evaluación de las destrezas orales en lengua inglesa: tipología, diseño de test y criterios de validación.

END NOTES

Note 1. The author is aware of the limited number of people participating in the study but considers the human sample involved significant and representative enough for a pilot study like this, which basically aims to raise questions, enhance reflection on evaluation scores and preliminarily showing the state of the art of the topic discussed for the subsequent and improved design and implementation of further studies involving a larger human sample (currently under development).

Note 2. Standard deviation is a measure of the dispersion of a set of data from its mean. If the data points are further from the mean, there is higher deviation within the data set.

Note 3. The coefficient of variation is a measure of spread that describes the amount of variability relative to the mean, that is, it is a statistical measure of the dispersion of data points in a data series around the mean and it is useful for comparing the degree of variation from one data series to another. In this study, in order to calculate the coefficient of variation of qualitative measures, marks have been converted according to their mean mark as shown in Table 1.

Note 4. The only discrepancy is that Peer mean QN score is slightly higher (6.58) than Teacher mean QL score (6.4)

Note 5. Coincidences are established according to the band correspondence for numerical QL evaluation interpretation in Table 1.

REFERENCES

Boud, D. and Lublin, J. (1983) Student self assessment: educational benefits within existing resources. In *Innovation through Recession*, ed. Geoffrey Squires (ed.) Guilford, Surrey: Society for Research into Higher Education, 93-99.

Boud, D. and Middleton, H. (2003) Learning from others at work: Communities of practice and informal learning. In *Journal of Workplace Learning, 15*(5), 194-202.

Carey, M., Mannell, R.H. and Dunn, P.K. (2011) Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201–219.

Cheng, W. and Warren, W. (2005). Peer assessment of language proficiency. In *Language Testing* 22 (1): 93-121.

Chen, L. and Jang. J.R. (2012) Improvement in automatic pronunciation scoring using additional basic scores and learning to rank. In *Proceedings of INTERSPEECH 2012*, Portland, Oregon.

Chen, L. and Jang. J.R. (2015). Automatic Pronunciation Scoring with Score Combination by Learning to Rank and Class-Normalized DP-Based Quantization. In *IEEE Transactions on Audio, Speech, and Language Processing, 1*(23), 1737-1749.

Cincared, T., Gruhn, R., Hacker, C., Nöth, E., and Nakamura, S. (2009) Automatic pronunciation scoring of words and sentences independent from the non-native's first language. In *Computer Speech and Language, 23*(1): 65–88.

Cucchiarini, C., Strik, H., and Boves, L. (2000) Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. In *Speech Communication,* 30(2-3), 109–119.

Davies, A. (2003) *The Native Speaker: Myth and Reality*. Clevendon: Multilingual Matters.

Eckes, T. (2015) *Introduction to Many-Facet Rasch Measurement. Analyzing and Evaluating Rater-Mediated Assessments*. 2nd Revised and Updated Edition. Series: Language Testing and Assessment. Volume 22. Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien: Peter Lang.

Falchikov, N., and Goldfinch, J. (2000) Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks. In *Review of Educational Research*, *70*(3),287-322.

Franco, H., Neumeyer, L., Kim, Y. and Ronen, O. (1997) Automatic pronunciation scoring for language instruction. Proceedings International Conference on Acoustic, Speech and Signal Processing, 97, 1471-1474.

Franco, H., Neumeyer, L., Digalakis, V. and Ronen, O. (2000) Combination of machine scores for automatic grading of pronunciation quality. In *Speech Communication, 30*(2-3), 121–130.

Frankland, S. (ed.) 2007. Teaching with Group Work, Peer and Self Assessment. In *Enhancing Teaching and Learning through Assessment: Deriving an Appropriate Model*, ed. Steve Frankland, 143-195. Netherlands: Springer.

Hacker, C., Batliner, A., Steidl, S., Nöth, E. Niemann, H. and Cincarek, T. (2005) Assessment of Non-Native Children's Pronunciation: Human Marking and Automatic Scoring. In Proceedings of the *10th International Conference on SPEECH and COMPUTER (SPECOM 2005)*, ed. George Kokkinakis, Nikos Fakotakis, Evangelos Dermatas, Rodmonga Potapova, 123–126, University of Patras, Moskow State Linguistics University.

IELTS Speaking band descriptors (public version) https://takeielts.britishcouncil.org/sites/default/files/IELTS_Speaking_band_descriptors.pdf

Johnson, K. (1996) Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics, 34,* 485–499.

Jonsson, A. and Svingby, G. (2007) The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144.

Klenowski, V. (1995). Student self-evaluation processes in student-centred teaching and learning contexts of Australia and England. *Assessment in Education: Principles, Policy & Practice*, *2*(2), 145–154.

Langan, M., Shuker, D., Cullen, R., Penney, D. Preziosi, R. and Wheater, P. (2008). Relationships between student characteristics and self, peer and tutor evaluations of oral presentations. *Assessment & Evaluation in Higher Education 33*(2), 179-190.

Li, H., Xiong, Y., Zang, X., Kornhaber, M.L., Lyu, Y., Chung, K.S. and Suen, H.K. (2016). Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245-264.

Magin, D. and Churches, A. (1989) Using self and peer assessment in teaching design. In *Proceedings,World conference on engineering education for advancing technology*. Australia: Institution of Engineers, *89*(1), 640-644.

Miller, P.J. (2003). The effect of scoring criteria specificity on peer and self-assessment. *Assessment and Evaluation in Higher Education, 28*(4), 383–94.

Moustroufas, N. and Digalakis, V. (2007) Automatic pronunciation assessment of foreign speakers using unknown text. *Computer Speech and Language*, *21*(1), 219–230.

Neumeyer, L., Franco, H., Digalakis, V. and Weintraub, M. (2000) Automatic scoring of pronunciation quality. In *Speech Communication, 30*(2-3), 83–93.

Ross, J.A. (2006) The reliability, validity, and utility of self-assessment. *Practical Assessment, Research, and Assessment*, *11*(10), 1-13.

Smith, H., Cooper, A. and Lancaster, L. (2002) Improving the quality of undergraduate peer assessment: a case for student and staff development. *Innovations in Education and Training International, 39*(1), 71-81.

Stefani, L.A.J. (1994) Peer, self and tutor assessment: Relative reliabilities. In *Studies in Higher Education 19(1):* 69-75.

Stevens, K.N. (2002) Toward a Model for Lexical Access Based on Acoustic Landmarks and Distinctive Features, *Journal of the Acoustical Society of America, 111*, 1872-1891.

Strik, H., Khiet Truong, Febe de Wet and Catia Cucchiarini. (2009) Comparing different approaches for automatic pronunciation error detection. In Speech Communication, *51*(10): 845–852.

Taylor, L. (2006). The changing landscape of English: implications for language assessment. *ELTJ*60 (1), 51-60.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249-276.

Witt, S. and Young, S. (2000) Phone-level pronunciation scoring and assessment for interactive language learning. In *Speech Communication, 30*(2-3), 95–108.

APPENDIX I

General Instructions and Final Results Sheet (Student Version)

Please read attentively:

1. Listen to your own recording and fill in one of the copies of the "Quantitative assessment rubric or numerical rubric".
2. Listen to your recording again (if necessary) and read the "Qualitative assessment rubric or band rubric" carefully. Then place yourself in one of the bands.
3. Listen to your partner's recording and fill in the other copy of the "Quantitative assessment rubric or numerical rubric".
4. Listen to your partner's recording again (if necessary), read the "Qualitative assessment rubric or band rubric" carefully again (if necessary) and place your partner in one of the bands.
5. Calculate the arithmetic means obtained when evaluating yourself in the "Quantitative assessment rubric or numerical rubric". Include the number obtained in section A of the table entitled "SUMMARY OF FINAL RESULTS".
6. Indicate the most frequent band letter when evaluating yourself in the "Quantitative assessment rubric or numerical rubric". Also include, according to the "Band correspondence" table below, the numerical value of the band letter chosen. Include both the letter and number obtained (separated by /) in section B of the table "SUMMARY OF FINAL RESULTS".
7. Calculate the arithmetic means obtained when evaluating your partner in the "Quantitative assessment rubric or numerical rubric". Include the number obtained in section C of the table below "SUMMARY OF FINAL RESULTS".
8. Indicate the most frequent band letter when evaluating your partner in the "Qualitative assessment rubric or band rubric". Also include, according to the "Band correspondence" below, the numerical value of the band letter chosen. Include the letter and number obtained (separated by /) in section D of the table entitled "SUMMARY OF FINAL RESULTS".

Band Correspondence

A	9
B	8
C	7
D	6
E	5
F	4
G	3
H	2
I	1

SUMMARY OF FINAL RESULTS

Your name (person assessing):________________________________

Candidate's code (as assigned by the researcher):______________________________

Self-assessment		Peer-assessment	
A) Numerical rubric (arithmetic mean)	B) Band rubric (A-I/x)	C) Numerical rubric (arithmetic mean)	D) Band rubric (A-I/x)

General Instructions and Final Results Sheet (Teacher Version)

Please read attentively:

If possible, evaluate students one at a time. When you finish with candidate 1, then start with candidate 2, and so on.

1. Listen to the student's recording and fill in a copy of "Quantitative assessment rubric or numerical rubric".
2. Listen to the student's recording again (if necessary), read the "Qualitative assessment rubric or band rubric" carefully and place the candidate in the corresponding band.
3. Calculate the arithmetic means obtained when evaluating the student in the "Quantitative assessment rubric or numerical rubric". Include the number obtained in the column entitled "numerical rubric (arithmetic mean)" of the table below entitled "SUMMARY OF FINAL RESULTS".
4. Indicate the most frequent band letter when evaluating the student in the "Qualitative assessment rubric or band rubric". Also calculate, according to the "Band correspondence" table below, the numerical value of the band letter chosen. Include the letter and number obtained in the column entitled "Band rubric (A-I/X)" of the table below entitled "SUMMARY OF FINAL RESULTS".

Band Correspondence

A	9
B	8
C	7
D	6
E	5
F	4
G	3
H	2
I	1

SUMMARY OF FINAL RESULTS

Student code	**Numerical rubric (arithmetic mean)**	**Band rubric (A-I/x)**
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		

APPENDIX II

Quantitative assessment rubric or numerical rubric

DESCRIPTORS **How would you describe…**	**1** **NULL**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9** **PERFECT**	**COMMENTS**
1. Phonemes/sounds production (Are there any problems with sounds (vowels, consonants and diphthongs) and their combination which negatively affects intelligibility?)										
2. Intonation: aspects related to tone or pitch variation. (Does tone rise and fall in the appropriate places? Or does it sound monotone?)										
3. Clarity and intelligibility: the extent to which a listener actually understands an utterance or is able to decode a message										
4. Word stress: stress patterns in individual words. (Does stress fall on the appropriate syllable?)										
5. Rhythm: timing (stress timing, syllable timing, rhythm in sentences) and linking of words in connected speech (how each segment of speech is liable to influence the segments that surround it). (Does the speaker speak in a natural rhythm? Or does language sound abrupt or choppy?)										
6. Sentence stress: pattern of stress given to words arranged in a sentence, often serving to express emphasis, attitude, etc., Related to focus and special emphasis (prominence)										
7. Accent: related to accentedness or how 'a listener's perception of how a speaker's accent is different from that of the L1 community'										
8. Chunking: the candidate pauses in the right place										
9. Delivery: speech rate and loudness. (Does the speaker speak too loudly or quietly, too fast, or too slow?)										

Qualitative assessment rubric or band rubric
Adapted from: http://ielts-yasi.englishlab.net/DETAILED_BAND_SCORE_DESCRIPTORS.htm

Your band	**Peer band**	**Main pronunciation features and expanded descriptors**
A	A	Expert user Has fully operational command of the language: appropriate, accurate and fluent with complete understanding. The candidate sounds like a native English speaker, or like someone who speaks "international English". Possibly he has a slight hint of his/her native accent. In general, the candidate shows all the pronunciation features of a native English speaker with the only errors being the kind of errors that even reasonably well educated native English speakers might occasionally make.
B	B	Very good user Has fully operational command of the language with only occasional unsystematic inaccuracies and inappropriacies. Misunderstandings may occur in unfamiliar situations. Handles complex detailed argumentation well.

(*Contd...*)

(*Continud*)

Your band	**Peer band**	**Main pronunciation features and expanded descriptors**
		Accent The examiner might be able to recognise the candidate›s native accent but it is (usually but not always) slight and does not interfere with the English pronunciation in any way. Understandability The pronunciation is very clear and accurate. No need to listen to any chunk twice. Basic Sound Accuracy All the vowels, consonants and diphthongs are pronounced very accurately, the way a native speaker pronounces them. Sentence stress Almost all the time, the candidate places the sentence stress on the correct word, although there might be one or two times when the examiner feels the wrong word was stressed. To a greater degree than “good users” (band below), this candidate makes use of sentence stress to express meaning or for emphasis. Intonation The candidate frequently shows the ability to vary his intonation to express meaning. Speed of delivery and “Chunking” The candidate shows the ability to consistently vary his/her speed by speaking in “chunks” of word groups. Word stress The candidate might make a rare error in placing the stress in a multi-syllable word on the wrong syllable but these errors would be in less common words and they would be the type of error that some native English speakers might make.
C	C	Good user Has operational command of the language though with occasional inaccuracies, inappropriacies and misunderstandings in some situations. Generally handles complex language well and understands detailed reasoning. Accent The candidate’s native accent might still be recognisable, but the accent of a native English-speaker is more dominant than the native accent. Understandability The pronunciation is very clear and accurate almost all the time. The examiner never or rarely needs to ask the candidate to repeat anything. There might be one or two times when the examiner needs to “think twice” about what word the candidate just said but it is rare for the examiner to need to ask the candidate to repeat his/herself. There should be no “patches” of language or short combinations of words that the examiner does not understand at all. However, the examiner might have to quickly “think twice” about the meaning of one or two “patches of language”. Basic sound accuracy Except for perhaps one or two occasions, the candidate pronounces all letters and diphthongs accurately. Sentence stress Most of the time, the candidate accurately places the sentence stress on the correct word in order to accurately express or emphasise his/her meaning. The candidate might make one or two errors when placing the word stress in noun+noun or adjective+noun combinations but these errors are usually for the lesser-known word combinations. Intonation The candidate shows good knowledge of how native English-speakers use intonation, i.e., a rising or falling tone, to communicate meaning. Speed of delivery and “Chunking” The candidate frequently shows some knowledge of how native English-speakers vary their speech speed to show meaning. This is mostly shown by his/her ability to speak in “chunks” of “word groups” faster than his/her parts of his/her sentences. However, he/he/she probably does not consistently speak in “chunks”. Linked sounds As part of his/her ability to show the skill of “chunking”, the candidate shows good ability at linking his/her speech sounds so that these chunks can be spoken faster. These chunks are spoken almost as if they were one long word. Along with an overall ability to speak linked sounds, the candidate shows a few instances of being able to link and blend his/her sounds very much like a native speaker does (or some native speakers do), for some short word combinations. For example, he/he/she might say, “dIdʒ’ə” for, “did you”.

(*Contd...*)

(***Continud***)

Your band	Peer band	Main pronunciation features and expanded descriptors
		Word stress There might be one or two instances of the candidate mispronouncing a multi-syllable word by stressing the wrong syllable but these pronunciation mistakes do not cause confusion.
D	D	Competent user Has generally effective command of the language despite some inaccuracies, inappropriacies and misunderstandings. Can use fairly complex language, particularly in familiar situations.
		Accent The candidate's native accent is possibly quite obvious but it does not interfere with understandability. At the same time, the national accent of an English-speaking country might also be discernible. Understandability The candidate speaks clearly most of the time but there might be 2 or 3 times when the examiner does not understand a word and needs to ask the candidate to repeat what he/she just said. Perhaps once or twice the examiner does not understand a "patch" of language such as a short combination of words. Basic sound accuracy There might be occasional inaccuracy in a few of the vowel, consonant and diphthong sounds but the examiner usually can guess what word the candidate is saying. For example, saying "ship" instead of "sheep". Overall, the candidate does not habitually mispronounce any one vowel, consonant or diphthong, although he/she might randomly mispronounce some of these at times. Sentence stress Overall, he/she shows some knowledge of correct sentence stress, although he/she does not use correct sentence stress in every utterance. The candidate has some knowledge about which word to stress in noun+noun combinations or adjective+noun combinations but might still make a few mistakes with this. The candidate shows some knowledge of how to stress keys words in a sentence for emphasis, such as when contrasting. Intonation The candidate shows some knowledge of how to use intonation, for example, when speaking a list, but is not consistent with his/her use of correct intonation. Linked speech sounds The candidate mostly links his/her speech sounds in a natural way but occasionally speaks each word separately, like a robot. Word stress The candidate might stress the wrong syllable in a multi-syllable word a small number of times but the word is usually recognisable to the examiner.
E	E	Modest user Has partial command of the language, coping with overall meaning in most situations, though is likely to make many mistakes. Should be able to handle basic communication in own field.
		Overall If the examiner feels that the candidate's pronunciation is better than the band below but not quite as good as the band above, the score is this band. Accent The candidate's native accent might be strongly evident and may, at times, result in mispronunciation of some sounds. However, the examiner is often able to guess the meaning when the candidate mispronounces a vowel, consonant or a diphthong. Understandability The candidate speaks clearly most of the time but there are about 4 or 5 times in the test when the examiner doesn't understand the pronunciation of a single word. There also might be about 3 or 4 times in the test when the examiner doesn't understand a "patch" of language, such as a part of a sentence or a complete, short sentence. Basic sound accuracy There is inaccuracy in a few of the vowel, consonant and diphthong sounds but usually the examiner can guess what word the candidate is saying. For example, saying "ship" instead of "sheep" and saying "maths" so that it sounds like "mice". The candidate might habitually mispronounce one or more of the vowel, consonant or diphthong sounds. Intonation The candidate speaks with a natural rising and falling tone at times but at other times speaks in a "flat" or "wooden" monotone.

(***Contd...***)

(*Continud*)

Your band	**Peer band**	**Main pronunciation features and expanded descriptors**
		Sentence stress He/she has some knowledge about stressing the key word in a sentence, for example, when speaking about contrasts, but only sometimes does this correctly. The candidate has little understanding of which word to stress in noun+noun combinations or adjective+noun combinations. When speaking these combinations, he/she is more or less guessing which word to stress and, in an attempt to avoid an error, often stresses neither word. Linked speech The candidate frequently doesn't link his/her speech sounds or words and, instead, frequently speaks each word separately, like a robot. Word stress The candidate stresses the wrong syllable in a multi-syllable word a few times but usually the word is recognisable to the examiner.
F	F	Limited user Basic competence is limited to familiar situations. Have frequent problems in understanding and expression. Is not able to use complex language. Accent The candidate's native accent is so strong that it interferes to a large extent in his/her English pronunciation. Understandability The examiner can understand what the candidate is saying about 70% of the time – the other 30% is unintelligible or very difficult to understand. Intonation and sentence stress The candidate might occasionally show some examples of correct rising/falling intonation and stressing the correct word in a sentence but mostly speaks in a monotonic way, like a robot. He/she might sometimes attempt to stress one particular word in a sentence but he/she lacks the understanding of which word to stress, with the result that these attempts are usually random guesses at which word to stress. Word stress The candidate stresses the incorrect syllable in a multi-syllable word several times.
G	G	Extremely limited user Conveys and understands only general meaning in very familiar situations. Frequent breakdowns in communication occur. Accent The candidate has a very heavy native accent that severely interferes with his/her English pronunciation. Understandability The examiner can only understand the candidate's pronunciation less than 50% of the time. Much of what the candidate says is unintelligible or very difficult to understand. Virtually everything the candidate says is spoken in a monotonic way, like a robot.
H	H	Intermittent User No real communication is possible except for the most basic information using isolated words or short formulae in familiar situations and to meet immediate needs. Has great difficulty in understanding spoken and written English. The examiner can only recognise a few English words in what the candidate says, and the candidate usually says very little.
I	I	Non user Essentially has no ability to use the language beyond possibly a few isolated words. The examiner can hardly recognise that the candidate is speaking English. This candidate speaks almost nothing, anyway.

APPENDIX III

Tasks to Record on the Part of Students (versions A, B and C)

A

Choose 1 of the following generalisations and talk about it for 2 minutes (you may either agree or disagree with it):

- Married people are boring.
- Footballers are not intelligent.
- You can't be friends with your boss.

Read the following text aloud:

Coronation Street

Coronation Street is Britain's longest-running television soap opera, and the UK's consistently highest-rated show. It was created by Tony Warren and first broadcast on the ITV network on Friday December 9, 1960. The working title of the show was Florizel Street, but Agnes, a tea lady at Granada Television, Manchester, (where Coronation Street is produced) remarked that "Florizel" sounded too much like a disinfectant. Jubilee Street was another option considered.

Coronation Street (commonly nicknamed Corrie, and also Coro St, Corra or even Corruption Street) is set in a fictional street in the fictional industrial town of Weatherfield which is based on Salford, now part of Greater Manchester (a Coronation Street does exist in Salford). Its principal rival soap operas are ITV1's Emmerdale and BBC1's EastEnders.

The show's iconic theme music, a brass-band throwback to the sounds of the 1940s, was written by Eric Spear and has been only slightly modified since the show's beginning.

Coronation Street can be seen on ITV1 on Sunday, Monday, Wednesday and Friday at 7:30 p.m. There is also an extra episode on Monday night at 8:30 p.m.

Granada and ITV executives, as well as the people in charge of distributing the show overseas, have called (and still call, as of 2006) Coronation Street the world's longest-running soap opera. The Guinness Book of Records recognises American soap opera Guiding Light as the world's longest-running soap opera, with over fifty years on television and an extra fifteen on radio.

From: http://www.saberingles.com.ar/reading/coronation-street.html

Read the following sentences aloud:

1. You've progressed well this year, but I'd like to see more progress.
2. In the desert, there is a big contrast between temperatures in the day and night.
3. Walter walked towards the waiter; Walter's waiter walked away.
4. The man controlled the nation's gold.
5. I saw the bird fly away.

B

Choose 1 of the following generalisations and talk about it for 2 minutes (you may either agree or disagree with it):

- Old people have no fun.
- Men are bad at languages.
- Young girls are brighter than young boys.

Read the following text aloud:

London Underground

The London Underground is a public transport network, composed of electrified railways (that is, a metro system) that run underground in tunnels in central London and above ground in the city's suburbs. The oldest metropolitan underground network in the world, first operating in 1863, the London Underground is usually referred to as either simply "the Underground" by Londoners, or (more familiarly) as "the Tube".

Since 2003, the Tube has been part of Transport for London (TfL), which also schedules and lets contracts for the famous red double-decker buses. Previously London Transport was the holding company for London Underground.

There are currently 275 open stations and over 253 miles (408 km) of active lines, with three million passenger journeys made each day (927 million journeys made 1999-2000; there are a number of stations and tunnels now closed).

Lines on the Underground can be classified into two types: sub-surface and deep level. The sub-surface lines were dug by the cut-and-cover method, with the tracks running about 5 metres below the surface. Trains on the sub-surface lines have the same loading gauge as British mainline trains.

The deep-level or "tube" lines, bored using a tunnelling shield, run about 20 metres below the surface (although this varies considerably), with each track running in a separate tunnel lined with cast-iron rings. These tunnels can have a diameter as small as 3.56m (11ft 8.25in) and the loading gauge is thus considerably smaller than on the sub-surface lines, though standard gauge track is used. […]

From: http://www.saberingles.com.ar/reading/underground.html

Read the following sentences aloud:
1. We import too much petrol and the country's export figures are going down.
2. Tim and Heather worked together; Heather never worked alone.
3. Susan ate six sweets at six o'clock and was sick.
4. The people queued to buy the food.
5. What time did the guests leave?

C

Choose 1 of the following generalisations and talk about it for 2 minutes (you may either agree or disagree with it):
1. All politicians are corrupt.
2. People who act are basically exhibitionists.
3.City people are more cultured than those from the country.

Read the following text aloud:
Red Telephone Box
The red telephone box, a public telephone kiosk designed by Sir Giles Gilbert Scott, was a once familiar sight on the streets of the United Kingdom. It has all but disappeared in recent years, replaced by a number of different designs. The few kiosks that remain have not been replaced because they are regarded as being of special architectural and historical interest.

The first standard public telephone kiosk introduced by the United Kingdom Post Office was produced by Somerville & Company in 1920 and was designated K1 (Kiosk no. 1). This design was not of the same family as the familiar red telephone boxes.

The red telephone box was the result of a competition in 1924 to design a new grander kiosk. The competition attracted designs from a number of noted architects. The Fine Arts Commission judged the competition and selected the design submitted by Sir Giles Gilbert Scott as the winner. The Post Office made a request that the material used for the design be changed from mild steel to cast iron, and that a slight modification be made to the door; after these changes, the design was designated K2. The kiosks were painted red was so that they might be easily recognised from a distance by a person in an emergency. In some rural areas the boxes were painted green so as not to disrupt the natural beauty of the surroundings.

From 1927 K2 was mainly deployed in and around London. K3 designed in 1930, again by Gilbert Scott was similar to K2 but was constructed from concrete and intended for rural areas. K4 (designed by the Post Office Engineering Department and proposed in 1923) incorporated a machine for buying postage stamps on the exterior. Only 50 kiosks of this design were built. […]
From: http://www.saberingles.com.ar/reading/red-telephone-box.html

Read the following sentences aloud:
1. It started as a student protest, but now the army has rebelled against the government.
2. These companies produce household objects such as fridges and washing machines.
3. Lenny talked a lot but he never talked to Lottie.
4. The points we scored are on the board.
5. He broke his arms in the accident.