Comparative Study on Speaking Assessment Rubrics in Trinity and Cambridge Language Certificates
Adaptation to the Common European Framework Guidelines

Lucia FRAGA-VIÑAS
Universidade da Coruña (A Coruña, Spain)
lucia.fragav@udc.es

Maria BOBADILLA-PEREZ
Universidade da Coruña (A Coruña, Spain)
m.bobadilla@udc.es

Abstract. The European Council has been instrumental in the standardization of language competence levels and certifications with the guidelines provided in the *Common European Framework of Reference for Languages* (CEFR) published in 2001 and later reviewed in 2020 with the *Companion Volume with New Descriptors* (CEFRCV). Cambridge Assessment English and Trinity College are two of the highest regarded institutions at the international level that grant their language certificates following the language competence levels provided by the CEFR. For this reason, the current study is grounded on the conviction that those certificates should meet certain principles of the Framework as a form of guarantee that they are assessing the CEFR level correctly. In particular, this paper focuses on the speaking skill and the rubrics of assessment used by the two afore-mentioned institutions. The rubrics of Trinity and Cambridge for the assessment of the oral production at the B2 CEFR level were considered for the purposes of this study – in particular, the rubrics that assess the oral production in the Integrated Skills in English (ISE-II) exam and in the First Certificate in English (B2 First). With a qualitative document research approach, this study analyses these rubrics in order to determine to what extent they respect the criteria established by the CEFR.

Keywords: CEFR, assessment rubrics, language certificates, speaking skill

1. Introduction

The field of teaching and learning of English as a Foreign Language in Europe has experienced great changes in the 21st century, mainly prompted by the establishment of the *Common European Framework of Reference for Languages: Teaching, Learning and Assessment* (CoE 2001). The strengthening of the European Union together with the market demands that globalization has imposed have brought about a greater focalization on the development of communicative skills. Before that, traditional foreign language lessons used to focus on learning vocabulary lists by heart and practising grammar with drilling and fill-in-the-blank type exercises. As a result, most learners were never fluent in the language.

The communicative approach promoted by the Council of Europe stirred up the existing methodologies. The aim is that the learner develops a good communicative competence. That is, being able to produce and receive oral and written text but also having knowledge of what is right and what is not in a concrete situation within a specific community of speakers. Moreover, it is essential that the learner develops some awareness of pragmatics, discourse and culture competences or some strategies to overcome possible difficulties.

In education, the implementation of the communicative approach has triggered the transformation of syllabi, tasks methodologies, and assessment processes. Besides establishing a scale of language levels of competence in a language, the CEFR works as a source of reference for the creation and design of education curricula and didactic materials, language certificates, and instruments of evaluation. The assistance in the creation of instruments of assessment, in particular rubrics (also called grading or rating scales), was indeed one of the aims of the CEFR. The Framework intended to be "a source for the development of rating scales for the assessment of the attainment of a particular learning objective and the descriptors may assist in the formulation of criteria" (CoE 2001: 179).

The CEFR also serves the purposes of establishing a background to "relate national and institutional Frameworks to each other" and "map the objectives of particular examinations and course modules using the categories and levels of the scales" included in the Framework (CoE 2001: 182). Thus, it established six levels of competence that range from users with a basic competence of the language (A1 and A2) through independent users (B1 and B2) to advanced users (C1 and C2). These levels have allowed the homogenization of official certificates, so an official B2 certificate obtained in Spain is supposed to be equal in level to any other official B2 certificate obtained in a different country.

The CEFR has been complemented and enhanced recently with the *Companion Volume with New Descriptors* (CEFRCV), which was published in 2020. In the Companion Volume, the notion of the learner/user as a social agent and the

action-oriented approach are expanded. Furthermore, the concepts of mediation and plurilingualism (North 2020a: 554) are further explored.

In the current globalized world and the demanding labour market, foreign language certificates become highly relevant because they favour the mobility of citizens among countries. Without any doubt, the CEFR has played a fundamental role in the standardization of those certificates with the establishment of levels of language competence, also called levels of proficiency. The CEFR is not a prescriptive document, and countries and language institutions can decide to what extent they implement or adapt the Framework attending to their specific contexts. Nevertheless, the institutions that grant official language certificates of one specific level of the CEFR all over the world should certainly use valid instruments of assessment and should at least respect the basic guidelines included in the CEFR regarding the design of grading scales.

This study intends to analyse two of the best-known and recognized English Certificates at Spanish and European level: Cambridge Assessment First Certificate (B2 FIRST) and Trinity College ISE-II, which certify the most commonly required level: upper-intermediate (B2). Thus, the research questions defined for this study are:

RQ 1 Do Cambridge B2 FIRST and Trinity College ISE-II certificates respect the criteria established by the CEFR?

RQ 2 Can some patterns be established regarding the type of assessment speaking rubrics used by the selected official English certificates?

2. Literature Review

2.1 The CEFR and the CEFRCV

As cited in the CEFR, assessment is used in the document "in the sense of the assessment of the proficiency of the language user" (CoE 2001: 177). The CEFR provides the entire educational community with learning standards for the teaching of foreign languages. In addition, teachers can construct the specifications of a task or test items using the document as a valuable source of reference. For those purposes, users of the CEFR (2001) can find information related to task specifications in section 4.1: "the context of language use" (domains, conditions and constraints, mental context), section 4.6: "Texts", and Chapter 7: "Tasks and their role in language teaching". Concerning the construction of test items, information can be found in section 5.2: "Communicative language competences" (CoE 2001).

The CEFR can also provide learning standards and guidelines for the construction of tasks. The descriptors (short texts that contain a description of what

each level of reference consists of) can be used both by teachers for assessment and by students for self-assessment. On the one hand, the descriptors for communicative acts, for instance, may be particularly helpful for giving feedback. Students get an overall impression of their performance in a task just by reading them. Scales may also be a good tool for summative assessment since teachers can build their rubrics or checklists on the grounds of the CEFR. The huge number of descriptors provided and classified according to their level are a great source for the development of rating scales and checklists. Furthermore, teachers can self-assess themselves or use the scale to implement student self-assessment. For instance, they can create a checklist or a type of grid for continuous assessment or for summative assessment at the end of each lesson/unit or course. Finally, the scales of the CEFR levels aim at enabling comparison among systems. In this regard, if the same descriptors are used in the examination, different tests can be compared as well as the results of those tests, so both national and institutional systems can be related.

As it has been mentioned, CEFR is very useful for the creation of rubrics – Chapter 9 gives guidelines on the construction of rubrics so that they can be feasible tools of assessment. This feasibility means that the teacher must be able to accurately assess all the criteria included in the rubric. With regard to this, the document emphasizes that "more than 4 or 5 categories starts to cause cognitive overload and that 7 categories is psychologically an upper limit" (CoE 2001: 193). This means that if a rubric measures 10 aspects or criteria, it is impossible for the teacher to assess each of them accurately with all the students. As a result, the CEFR recommends that in the event the limit is exceeded, features should be combined and renamed under a broader category.

Appendix A of the CEFR includes several specifications for the formulation of descriptors in a rubric. The first remark is *positiveness*. Previous research on proficiency scales detected a tendency to formulate lower-level descriptors with negative sentences. The CEFR acknowledges the difficulty in doing so: "it is more difficult to formulate proficiency at low levels in terms of what the learner can do rather than in terms of what they can't do" (CoE 2001: 205), but it also encourages the desire to revert that tendency.

Definiteness in the statements is also encouraged. *Avoiding vagueness* and describing concrete tasks are essentials for achieving effectiveness. However, definiteness should not lead to the production of excessively long descriptors since, as the framework notes, a "descriptor which is longer than a two-clause sentence cannot realistically be referred to during the assessment process" (CoE 2001: 207). *Brevity* also helps the independence of descriptors. Moreover, the descriptors must be *clear and transparent* so that both the examiner and the learner can completely understand what is expected in the assessment.

The CEFRCV added new descriptors and modified some of those already existing in the original document, but especially it added a new descriptive

scheme with the definition and construction of mediation (Piccardo 2019: 6). Mediation implies that the teaching-learning process must be oriented towards learners. Thus, students must be able to perform in real-life situations. This is the reason why a focus on interaction and construction of meaning has been included in the CEFRCV (Foley 2019: 30). This construction of meaning may occur in different forms, as the learner can take it from other languages he/she knows or studies (translanguaging). The development of the concept of mediation is one of the fundamental key aspects of the CEFRCV (CoE 2020: 91). It is explained as follows: "the term mediation is also used to describe a social and cultural process of creating conditions for communication and co-operation, facing and hopefully defusing any delicate situations and tensions that may arise".

North et al. (2019: 21) define mediation as a process through which language is a vehicle to access other new concepts and that normally involves reception, production, and interaction. Nineteen scales for mediation activities and five scales concerning mediation strategies have been included in the CEFRCV. Among those, scales for explaining data in speech and writing, expressing personal response, analysis and criticism to creative texts, leading and collaborating in group work, facilitating communication in delicate situations and disagreement, or simplifying a text have been created, too, for the CEFRCV. Those scales are accompanied by suggestions of activities and tasks to work and/or assess mediation. The following is one of the new scales added:

EXPLAINING DATA IN SPEECH (E.G. IN GRAPHS, DIAGRAMS, CHARTS ETC.)	
C2	Can interpret and describe clearly and reliably (in Language B) various forms of empirical data and visually organised information (with text in Language A) from conceptually complex research concerning academic or professional topics.
C1	Can interpret and describe clearly and reliably (in Language B) the salient points and details contained in complex diagrams and other visually organised information (with text in Language A) on complex academic or professional topics.
B2	Can interpret and describe reliably (in Language B) detailed information contained in complex diagrams, charts and other visually organised information (with text in Language A) on topics in his/her fields of interest.
B1	Can interpret and describe (in Language B) detailed information in diagrams in his/her fields of interest (with text in Language A), even though lexical gaps may cause hesitation or imprecise formulation.
	Can interpret and describe (in Language B) overall trends shown in simple diagrams (e.g. graphs, bar charts) (with text in Language A), even though lexical limitations cause difficulty with formulation at times.
A2	Can interpret and describe (in Language B) simple visuals on familiar topics (e.g. a weather map, a basic flow chart) (with text in Language A), even though pauses, false starts and reformulation may be very evident in speech.
	No descriptors available
A1	*No descriptors available*
Pre-A1	*No descriptors available*

Source: CoE 2020: 109

Figure 1. *One of the scales included for mediation in the CEFRCV with New Descriptors*

Mediation is not the only novelty of the CEFRCV. Two scales for online interaction and goal-oriented transactions have also been incorporated together with scales to deal with literature and sign languages as well as a development of a phonological scale. The original scales have been maintained although some of the descriptors have been retouched or expanded, particularly those related to the C2 level, the A1, or de pre-A levels. For instance, any reference to "native speakers" has been erased from the descriptors, as it is understood that C2 level is not equivalent to native level. In addition, the description for plus levels (B1+/ B1.2) has been strengthened. The following figure illustrates some of the changes. It shows one of the scales that was already part of the original document (parts in blue) combined with the new modifications (written in black):

SUSTAINED MONOLOGUE: DESCRIBING EXPERIENCE		PROSIGN
C2	Can give clear, smoothly flowing, elaborate and often memorable descriptions.	
C1	Can give clear, detailed descriptions of complex subjects. Can give elaborate descriptions and narratives, integrating sub themes, developing particular points and rounding off with an appropriate conclusion.	
B2	Can give clear, detailed descriptions on a wide range of subjects related to his/her field of interest. Can describe the personal significance of events and experiences in detail.	
B1	Can clearly express feelings about something experienced and give reasons to explain those feelings. Can give straightforward descriptions on a variety of familiar subjects within his field of interest. Can reasonably fluently relate a straightforward narrative or description as a linear sequence of points. Can give detailed accounts of experiences, describing feelings and reactions. Can relate details of unpredictable occurrences, e.g. an accident. Can relate the plot of a book or film and describe his/her reactions. Can describe dreams, hopes and ambitions. Can describe events, real or imagined. Can narrate a story.	
A2	Can tell a story or describe something in a simple list of points. Can describe everyday aspects of his/her environment e.g. people, places, a job or study experience. Can give short, basic descriptions of events and activities. Can describe plans and arrangements, habits and routines, past activities and personal experiences. Can use simple descriptive language to make brief statements about and compare objects and possessions. Can explain what he/she likes or dislikes about something. Can describe his/her family, living conditions, educational background, present or most recent job. Can describe people, places and possessions in simple terms. Can say what he/she is good at and not so good at (e.g. sports, games, skills, subjects). Can briefly talk about what he/she plans to do at the weekend or during the holidays.	
A1	Can describe him/herself, what he/she does and where he/she lives. Can describe simple aspects of his/her everyday life in a series of simple sentences, using simple words and basic phrases, provided he/she can prepare in advance.	
Pre-A1	Can describe him/herself (e.g. name, age, family), using simple words and formulaic expressions, provided he/she can prepare in advance. Can say how he/she is feeling using simple words like 'happy', 'tired', accompanied by body language.	

Source: CoE 2018: 70

Figure 2. *Scale for speaking production included in the CFRCV with new descriptors*

The original CEFR, published in 2001, has been the object of multiple criticisms (Martyniuk–Noijons 2007, Bärenfanger et al. 2019, Deygers 2021). Most of the critical comments deal with the formulation of descriptors, their scope, the development and validation of descriptors or the density of the document itself. Brian North, co-author of the CEFR and the CEFRCV, addressed most of the criticism (2020b) arguing that the scales provided in the document were intended to be illustrative. Nevertheless, there are still authors with some concerns on the issue. They believe that if the arguments provided by the authors of the CEFR were consistent, the new CEFRCV would have not focused more than the original one on the scales, neither had it provided a closer look at their methodology (Foley 2019: 32). Notwithstanding, the impact of the CEFR on the teaching and learning of foreign languages worldwide is rather undeniable.

2.2. Speaking skill and rubrics

Productive skills involve the production of a text, either written or spoken. Currently, they are often assessed with a rubric. Before the publication of the CEFR, traditional methodologies, such as the Grammar–Translation Method, were still very common, so the spoken production did not use to be assessed. During the session, the student had to memorize vocabulary (frequently out of context) and had to translate texts. As a result, learners barely spoke in the foreign language. Luckily, the scenario has changed, and now the CEFR has promoted sessions grounded on the communicative competence.

The teaching and practice of the speaking skill in the classroom brings along quite a few changes in the evaluation system. In any oral exam, there are many different factors to bear in mind: for instance, the manner of articulation of the message or the body language. Bygate (in Baitman–Beliz 2012) states that the oral process consists of three main phases: conceptualization of the message content, linguistic formulation of the message, and articulation of the message. According to the same author, the evaluation of the speaking skill in L2 is particularly complex because it is time-consuming and it is very difficult to build up a real-life situation (in Baitman–Beliz 2012).

Rubrics or grading scales are charts that allow the assessment of a task on the grounds of some criteria and an established scale. The use of rubrics for the assessment of performance became a common practice in the late 20th and the early 21st century, particularly so with the publication of the CEFR. This is so because, as it was explained, traditional methodologies used in language classrooms up to the 80s did not focus as much on performance as on language learning with a main emphasis on vocabulary and grammar practice (Bobadilla-Pérez–Fraga-Viñas 2020: 162). A rubric normally consists of a grid with multiple cells. Rubrics may have only two sections (if they are holistic) or four (when

they are analytic). When the rubric is composed of two main sections, there is one vertical column and a horizontal one as well. One of them corresponds to the language descriptors and the other one to the scores. The two sections are frequently assembled, and they form just one column in which each cell contains the score and the language description of that level.

Rubrics that have four sections normally contain a task description at the top. They usually have many columns, rows, and cells. The first horizontal row is usually the scale. The scale levels may be numbers, but they are often words that indicate the level of achievement (Excellent, Good, Poor, etc.). The first column on the left will contain the dimensions, also called criteria or categories. These will indicate the aspects that are being measured (Grammar, Vocabulary, Cohesion, Coherence, Fluency, Clarity, Visual Contact, etc.). Finally, the remaining cells in the grid are the descriptors of the performance for each criterion and level of the scale.

There is also a large variety of rubric types. The most common criterion to classify rubrics is according to how performance is assessed: as a whole or separately. If the assessment of the performance is divided into criteria (grammar, cohesion, content, etc.), it is an analytic rubric, and when the performance is assessed as a whole, it is holistic. Rubrics can also be classified according to the scale they used. Hence, rubrics can be quantitative (numerical scale), qualitative (nominal scale), or mixed (when they use both numerical and nominal scale). Rubrics can also be hyper-generalized, when they can be used to assess any skill, skill-focused, when they only assess one particular skill (e.g. speaking), or task-focused, when they are tailored to one particular task (e.g. rubric for the assessment of a speaking presentation). Finally, rubrics can also be distinguished by their function. Thus, when they are used to check to what extent the learner has acquired the content taught, they are called achievement rubrics, and when they are used to determine the learner's level, they are proficiency rubrics.

Surprisingly, there are not many studies on speaking rubrics. Among them, Luu Tron Tuan's (2012) research is worth mentioning. The study involved 104 students divided in two groups, and it was related to the use of holistic and analytic rubrics to assess speaking. The research consisted in examining students six times and allowing them to see their results after each of the examinations. It concluded that those students that had been assessed with an analytic rubric improved their qualification more than one point, while those who had been assessed with a holistic rubric barely improved.

The assessment of the speaking skill in general has been the object of much research. The one carried out by Brittany Baitman and Mauricio Veliz (2012) or the one conducted by Emrah Ekmekçi (2016) could be highlighted, as they refer specifically to foreign languages. Ekmekçi's research involved eighty students of English and six teachers (three native speakers and three non-native). The study concluded that the differences found between native and non-native raters were

not statistically significant. However, Baitman and Veliz's study, which consisted of twelve teachers who assessed four TOELF tasks, proved that non-native teachers tend to give lower scores (2012: 186). Moreover, this research concluded that native teachers attach more importance to fluency, pronunciation, and vocabulary, while non-natives teachers care more about grammar and vocabulary (Baitman–Veliz 2012: 191).

When assessing speaking, it is also essential to think of the elements that are going to be taken into consideration. Extensive research on rubrics found out which criteria are the most frequent. Vocabulary range, grammar control, organization and structure, cohesive devices, content and pronunciation were the most common assessment criteria, as they appeared in 95% of the rubrics analysed (Schreiber et al. 2012). Many of these rubrics also assessed performance, body language, visual contact, and visual support.

3. Objective and methodology

3.1 Objective

The objective of the present study is to determine to what extent the speaking rubrics of Cambridge B2 First and Trinity ISE II language certificates follow the guidelines provided by the CEFR in the assessment of their speaking tests. The two institutions officially certify CEFR levels, so they should at least follow the basic guidelines of the CEFR regarding the creation of rubrics.

3.2. Method

The Council of Europe established in 1971 that teaching must be planned as a coherent whole and must cover objectives, evaluation achievements, and effectiveness. This requires a coordinated effort between administrations, materials, examiners, teachers, and students that should always share objectives and criteria (Vez 2011: 89). The current research uses a qualitative approach and the technique of document analysis. Moreover, the implementation of this qualitative approach was made following the seven steps of Gil Pascual's research methodology. The first step was the establishment of objectives and context. As it has been stated before, the aim was the study of rubrics for the assessment of speaking in official English Certificates to check whether they follow the guidelines established by the CEFR.

The second step was the definition of the units of analysis, that is, the selection of the rubrics. There was obviously a wide range of official English certificates, but only European institutions were considered, taking into account that the

CEFR was created by the Council of Europe. The selected examination centres were Cambridge Assessment English and Trinity College. The level chosen was B2, as it is the upper-intermediate and the one most frequently required to work and study in Europe. The CEFR establishes that at this level a user can interact with fluency and spontaneity and that the interaction with a native speaker is possible without great difficulty for any of the parties (CoE 2001: 24).

B2	Grammar and Vocabulary	Discourse Management	Pronunciation	Interactive Communication
5	Shows a good degree of control of a range of simple and some complex grammatical forms. Uses a range of appropriate vocabulary to give and exchange views on a wide range of familiar topics.	Produces extended stretches of language with very little hesitation. Contributions are relevant and there is a clear organisation of ideas. Uses a range of cohesive devices and discourse markers.	Is intelligible. Intonation is appropriate. Sentence and word stress is accurately placed. Individual sounds are articulated clearly.	Initiates and responds appropriately, linking contributions to those of other speakers. Maintains and develops the interaction and negotiates towards an outcome.
4	*Performance shares features of Bands 3 and 5.*			
3	Shows a good degree of control of simple grammatical forms, and attempts some complex grammatical forms. Uses a range of appropriate vocabulary to give and exchange views on a range of familiar topics.	Produces extended stretches of language despite some hesitation. Contributions are relevant and there is very little repetition. Uses a range of cohesive devices.	Is intelligible. Intonation is generally appropriate. Sentence and word stress is generally accurately placed. Individual sounds are generally articulated clearly.	Initiates and responds appropriately. Maintains and develops the interaction and negotiates towards an outcome with very little support.
2	*Performance shares features of Bands 1 and 3.*			
1	Shows a good degree of control of simple grammatical forms. Uses a range of appropriate vocabulary when talking about everyday situations.	Produces responses which are extended beyond short phrases, despite hesitation. Contributions are mostly relevant, despite some repetition. Uses basic cohesive devices.	Is mostly intelligible, and has some control of phonological features at both utterance and word levels.	Initiates and responds appropriately. Keeps the interaction going with very little prompting and support.
0	*Performance below Band 1.*			

Source: Cambridge English Language Assessment (2016)

Figure 3. *Rubric for the Assessment of Speaking in B2 FIRST*

ISE II Speaking & Listening rating scales

ISE II Speaking & Listening rating scale

This rating scale is used by the examiner to make a judgement of the candidate's performance in the speaking exam (the Topic, Collaborative and Conversation tasks).

Score	Communicative effectiveness	Interactive listening	Language control	Delivery
	▸ Task fulfilment ▸ Appropriacy of contributions/turn-taking ▸ Repair strategies	▸ Comprehension and relevant response ▸ Level of understanding ▸ Speech rate of examiner interventions ▸ Speed and accuracy of response	▸ Range ▸ Accuracy/precision ▸ Effects of inaccuracies	▸ Intelligibility ▸ Lexical stress/intonation ▸ Fluency ▸ Effects on the listener
4	▸ Fulfils the task very well ▸ Initiates and responds with effective turn-taking ▸ Effectively maintains and develops the interaction ▸ Solves communication problems naturally, if any	▸ Understands all interventions on a first hearing ▸ Interprets examiner's aims and viewpoints accurately by making links with earlier information ▸ Makes immediate and relevant responses	▸ Uses a wide range of grammatical structures/ lexis flexibly to deal with topics at this level ▸ Consistently shows a high level of grammatical accuracy and lexical precision ▸ Errors do not impede communication	▸ Clearly intelligible ▸ Uses focal stress and intonation effectively ▸ Speaks promptly and fluently ▸ Requires no careful listening
3	▸ Fulfils the task appropriately ▸ Initiates and responds appropriately ▸ Maintains and develops the interaction appropriately (eg expanding and developing ideas, and showing understanding of what the examiner said) ▸ Deals with communication problems well	▸ Understands most interventions on a first hearing ▸ Interprets examiner's aims and viewpoints accurately ▸ Makes prompt and relevant response	▸ Uses an appropriate range of grammatical structures/ lexis to deal with topics at this level ▸ Shows a relatively high level of grammatical accuracy and lexical precision ▸ Errors do not impede communication	▸ Clearly intelligible despite some use of non-standard phonemes ▸ Uses focal stress and intonation appropriately ▸ Generally speaks promptly and fluently – occasionally affected by some hesitancy ▸ Requires almost no careful listening
2	▸ Fulfils the task acceptably with support ▸ Initiates and responds acceptably ▸ Maintains and develops the interaction, but contributions are not always appropriate and/or somewhat dependent on the examiner ▸ Manages to solve communication problems, but requires more than one attempt and/or does not always do this naturally (eg 'What?')	▸ Usually understands interventions; occasionally needs clarification ▸ Shows occasional uncertainty about examiner's aims or viewpoints ▸ Makes relatively prompt responses	▸ Uses an acceptable range of grammatical structures/lexis to manage topics at this level, but grammatical/lexical gaps still cause hesitation and circumlocution ▸ Shows an acceptable level of grammatical accuracy and lexical precision ▸ Most errors do not impede communication	▸ Intelligible despite some use of non-standard phonemes ▸ Uses focal stress and intonation acceptably ▸ Speaks promptly and fluently enough to follow ▸ Requires some careful listening
1	▸ Does not fulfil the task even with support ▸ Does not initiate or respond adequately ▸ Does not maintain and develop the interaction sufficiently ▸ Contributions are inappropriate and/or overly dependent on the examiner ▸ Has some difficulty in resolving communication problems	▸ Has difficulty in understanding interventions ▸ Frequently misinterprets examiner's aims and viewpoints ▸ Responds slowly due to difficulty in understanding input	▸ Uses a limited range of grammatical structures/ lexis that is not always adequate to deal with topics at this level ▸ Does not show an adequate level of grammatical accuracy and lexical precision ▸ Some errors impede communication	▸ Generally intelligible or sometimes unintelligible. Use of non-standard phonemes is sometimes or frequently evident ▸ Sometimes or often misuses focal stress and intonation ▸ Speaks slowly. Sometimes or often halted by hesitancy ▸ Requires (some) careful listening
0	No performance to assess (candidate does not speak, or does not speak in English). Also use if no topic is prepared.			

Source: Trinity College London (2017)

Figure 4. *Rubric for one of the Tasks in the ISE-II*

The third step was the definition of the principles that were going to be used to determine whether the rubrics followed the CEFR guidelines. In this case, the CEFR establishes that rubrics, or grading scales must be feasible and should not have more than five criteria. Concerning descriptors, they should be written with positive statements, must be brief but precise, and vagueness in their formulation must be avoided. Furthermore, rubrics were also classified according to their typology, scale, topic, level of application, function, type of examiner, and format. Regarding the formulation of descriptors, the CEFRCV states that it is essential that "each is brief (up to 25 words), is clear and transparent, is positively formulated, describes something definite and has independent stand-alone integrity – not relying on the formulation of other descriptors for its interpretation" (CoE 2018: 41).

Categorization and codification made up the fourth step. Tables were selected as the instrument for data gathering. The visual construction was very handy to categorize data extracted from the rubrics. The instrument gathered information regarding the general typology of the rubric (with different variables such as measurement, scoring, theme, application, function, or channel), as well as the relevance and validity factors mentioned above. Lastly, it gathered specific information related to the CEFR and the CEFRCV descriptors with regard to their positive formulation, briefness, and vagueness.

Gil (2011) recommends the measurement of reliability. This can be carried out through a reliability coefficient such as Cronbach's alpha, and it refers to how consistent the results obtained with a rubric are. Finally, data were analysed as a whole and were compared so that contradictions with the CEFR guidelines, if any, could be detected and to check whether some updates are needed in order to match the new CEFRCV.

4. Results and discussion

Following the comparative analysis of the rubrics used for the assessment of speaking in Cambridge Assessment English B2 FIRST and in Trinity College ISE-II, it was found that both rubrics are analytic and not holistic. Analytic rubrics have been proved to be more reliable, as the research carried out by Tuan (2012) explained or as checked by other studies in the same line (Sundeen 2014, Becker 2016).

Both rubrics are quantitative and use a numerical scale. Nevertheless, the B2 FIRST rubric uses a scale from 0 to 5, whereas the ISE-II scale is shorter: from 0 to 4. None of them contains any quantitative scale together with the numerical one. The rubrics used in English certificates normally contain quantitative scales rather than qualitative ones because they want to be as accurate and precise as possible. Notwithstanding, if the number were accompanied by a quantitative

expression like "excellent for the level" or "acceptable for the level", as it is the case with the rubrics employed by the Spanish Official School of Languages, the feedback given to the student would be more complete.

According to the application, the ISE-II's speaking rubric is task-specific, which means that the rubric is especially built for the task being examined. In this case, the Trinity College's certificate assesses both the speaking and the listening skills in the same paper. This is the reason why there are two rubrics for the assessment of this part. One rubric (analytic) is for one task that involves speaking about a listening task, and there is also another rubric (holistic) for the assessment of an independent listening task. However, the rubric used in B2 FIRST is skill-focused, as the same rubric is used for the assessment of all the speaking tasks. Most of the English certificates use skill-focused rubrics to assess the speaking tasks (a monologue, describing a picture, interact in a discussion with a partner, etc.). However, in the ISE-II exam, the student has to listen to a track and, afterwards, summarize it using his/her own words so that the understanding of the audio is proven. Then, the examiner and the candidate hold a conversation that stems from the contents of the listening. This format allows for the assessment of the speaking and the listening skill together, as the new CEFRCV now encourages.

As far as the contrast with the CEFR guidelines for the creation of rubrics is concerned, the rubric used by Trinity College ISE-II consists of four criteria: communicative effectiveness, interactive listening, language control, and delivery – so, it could be a handy or feasible rubric in terms of criteria. As for the B2 FIRST's rubric, it assesses grammar and vocabulary, discourse management, pronunciation, and interactive communication – so, the total number of criteria, i.e. five, is also suitable.

Regarding the formulation of descriptors, the results of the analysis show that none of the rubrics really fulfils the recommendations either of the original CEFR or of the new CEFRCV. The rubric of the ISE-II contains negative descriptors, such as "does not maintain and develop interaction" although the attempt for using positive expressions is clear: "responds slowly", "occasionally needs clarifications", "uses a limited range of grammatical structures", "shows occasional uncertainty" (Trinity College London 2017). The B2 FIRST's rubric, on the other hand, does show positively worded descriptors, even at the lowest band: (band 1) "shows a good degree of control of simple grammatical forms", "uses basic cohesive devices", "keeps the interaction going with very little prompting and support" (Cambridge English Language Assessment 2016).

Regarding the vagueness or briefness of the descriptors, both rubrics have deficiencies. The rubric in B2 FIRST contains three vague bands of descriptors, e.g. "performance shares features of Bands 3 and 5" (Cambridge English Language Assessment 2016). This is obviously imprecise, and it also contradicts the new CEFRCV recommendations for the formulation of descriptors. Those

recommendations state that each descriptor should "[describe] something definite and has independent stand-alone integrity – not relying on the formulation of other descriptors for its interpretation" (Council of Europe 2018: 41). The rubric used in Trinity's certificate is not vague, and it does have proper descriptors for all the levels but for the lowest one (0). Notwithstanding, in terms of briefness, the ISE-II rubric descriptors are the opposite of brief, and they do not fulfil the "up to 25-word-long" established in the CEFRCV (2018: 41). This rubric contains descriptors that are almost double that figure (49 words). This makes the assessment of students with the rubric tedious and more time-consuming.

In connection with the changes brought about by the CEFRCV, it must be mentioned that both rubrics already include descriptors to measure phonetics and phonology that are equivalent to those included in the new phonological scale of the CEFRCV. Nevertheless, only the paper for the Trinity ISE-II incorporates somehow the concept of mediation, as the skills are assessed in a more integrated manner. To exemplify this, it can be explained that the speaking exam includes an oral summary of the contents of listening and includes a conversation directly brought about by those contents, so the speaking and the listening skills are being assessed at the same time in an integrated way. This is an aspect that has not been incorporated in Cambridge Certificate, which assesses skills in four separate papers: reading and use of English, writing, listening, and speaking.

Finally, it is also recommended that research based on the reliability and validity of the certificates should be published. Cambridge does publish the results of reliability coefficients of its papers: Cronbach's alpha (0.84) and SEM (1.50) for its speaking paper. These results point out a high reliability of the exam. However, the results provided date back to 2010, so it would be strongly advisable to update those studies. On the other hand, the Trinity College web contains a downloadable report about the validity of the certificate, with a high reliability figure (0.983) for the Cronbach's alpha of the speaking and listening part. Nevertheless, this report is from 2007, therefore its revision and update are desirable.

5. Conclusions and implications

The research carried out has yielded some interesting findings that allow some reflections. The study was intended to check the suitability of the rubrics used by the two official English certifications to the guidelines proposed by the CEFR. To begin with, it should be noted that rubrics are not always publicly available for those interested in obtaining official English certificates. In this sense, it is commendable that both institutions, Cambridge Assessment English and Trinity College, publish their assessment rubrics on their respective websites. In consequence, candidates can know what is expected from them. The study

conducted by Laurian and Fitzgerald in 2013 proved the importance of giving the students access to the rubric prior to the assessment process.

Concerning Research Question 2, the only patterns that can be established are that both certificates use proficiency rubrics and they both use analytic rubrics with numerical scales for the assessment of the speaking production. Nevertheless, while one is skill-focused, the other is task-focused. The number of criteria and the levels of the numerical scales are also different, being five in the case of the FIRST and four in the case of ISE-II. Although both certificates share high figures of reliability of these speaking tests, the figure for the Trinity College certificate is higher, which seems to indicate their assessment of speaking is a priori more reliable. However, it would be necessary to have more information about the conditions of the studies, the size of the samples, etc. as well as to update the results of that body of research in order to be able to do a meaningful interpretation of the results.

In answer to Research Question 1, it can be concluded that neither B2 FIRST nor ISE-II speaking rubrics completely follow the guidelines mentioned by the CEFR. Even though both rubrics have a feasible and manageable number of criteria, the descriptors for the rubric of ISE-II are not entirely written in positive sentences and they are not brief (they exceed the 25-word limitation). Meanwhile, the rubric used in B2 FIRST contains vague descriptors in the intermediate bands of levels. In addition, it also contradicts the guidelines incorporated in the new CEFRCV stating that descriptors should be completely independent from the ones in the level below and above (CoE 2018: 40).

A possible disadvantage of the CEFR or the CEFRCV is that the scales included in both documents are mainly holistic. North (2020b) argued that this is so because the scales provided are meant to be used only as a reference. However, it is undeniable that if an analytic scale must be designed taking as reference a holistic one, the process is much more complex. If the CEFR provided more analytic scales, the task of building a rubric tailored for one certificate would be speeded up and eased. There would also be many more examples of descriptors that are positively worded and brief but precise.

Finally, the CEFRCV with new descriptors has introduced some changes, but the most important one is the reinforcement and expansion of the concept of mediation. While some official certificates have already been updated to add tasks for the assessment of mediation in their papers, by the time this research had been conducted, neither the selected Cambridge Certificate nor the one of Trinity College has made any alterations to introduce mediation tasks, although the concept of mediation can be argued to be vaguely present in some of their current tasks (e.g. to speak about an image). Nevertheless, without any modification being made, the ISE-II tasks are more integrated since the different communicative language activities evaluated (reception, production) are assessed in pairs.

It is fundamental to restate that the CEFR and its new complementary version, the CEFRCV, are not prescriptive documents. This means that any country or institution can decide whether or not they want to implement their approach and to what extent. However, it seems coherent that if an institution is granting an official certificate of one CEFR level, it should follow the main and basic guidelines provided by the CEFR regarding assessment rubrics. To conclude, in the same way that a revision and update of the original CEFR was often demanded, any certificate granting CEFR-based levels should revise, improve, and update their exams, too, even if it is not mandatory. After all, revision and adjustment to new scenarios, realities, and problems are a constant in any learning process. Therefore, frameworks of reference, syllabi, materials, methods, and assessments should also be constantly under revision and adjustment in order to keep improving the learning process.

References

Baitman, Brittany–Mauricio Veliz. 2012. A comparison of oral evaluation ratings by native English teachers and non-native English speaker teachers. *Literatura y Lingüística* 61(3): 171–200. *https://scielo.conicyt.cl/scielo.php?script=sci_abstract&pid=S071658112013000100010&lng=e&nrm=iso* (Last accessed: 28 April 2022).

Bärenfänger, Olaf–Claudia Harsch–Bernd Tesch–Karin Vogt. 2019. Reform, Remake, Retousche? Diskussionspapier der Deutschen Gesellschaft für Fremdsprachenforschung zum Companion to the CEFR. *Zeitschrift für Fremdsprachenforschung* 30(1): 7–13. *https://dialnet.unirioja.es/ejemplar/519674* (Last accessed: 28 April 2022).

Becker, Anthony. 2016. Student-generated scoring: Examining their formative value for improving ESL students' writing performance. *Assessing Writing* 29: 15–24. *https://www.sciencedirect.com/science/article/pii/S1075293516300162* (Last accessed: 28 April 2022).

Bobadilla-Pérez, María–Lucia Fraga-Viñas. 2020. Making language learning more inclusive: Introducing rubrics to adult students to improve written performance. In Cristina-Aránzazu Huertas-Abril–Maria Elena Gómez-Parra (eds), *International Approaches to Bridging the Language Gap*, 152–167. IGI Global.

Brooks, Gavin. 2012. Assessment and academic writing: A look at the use of rubrics in the second language writing classroom. *Kwansei Gakuin University Humanities Review* 17: 227–240. *https://www.researchgate.net/profile/Gavin-Brooks-2/publication/324217649_Assessment_and_academic_writing_A_look_at_the_use_of_rubrics_in_the_second_language_writing_classroom/links/5ac584fca6fdcc051daf2a31/Assessment-and-academic-writing-A-look-*

at-the-use-of-rubrics-in-the-second-language-writing-classroom.pdf (Last accessed: 28 April 2022).

Cambridge English Language Assessment. 2016. *Cambridge English First Handbook for Teachers for Exams from 2016.* Cambridge: Cambridge English Language Assessment. *www.cambridgeenglish.org/images/167791-cambridge-english-first-handbook.pdf* (Last accessed: 29 April 2022).

Council of Europe (CoE). 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Strasbourg: Council of Europe.

—— 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors.* Strasbourg: Council of Europe.

—— 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors.* Strasbourg: Council of Europe.

Deygers, Bart. 2021. The CEFR companion volume: Between research-based policy and policy-based research. *Applied Linguistics* 42(1): 186–191. *https://academic.oup.com/applij/article/42/1/186/5487749?login=true* (Last accessed: 28 April 2022).

Ekmekçi, Emrah. 2016. Comparison of native and non-native English language teachers' evaluation of EFL learners' speaking skills: Conflicting or identical rating behaviour? *English Language Teaching* 9(5): 98–105. *https://files.eric. ed.gov/fulltext/EJ1097579.pdf* (Last accessed: 28 April 2022).

Foley, Joseph A. 2019. Issues on assessment using CEFR in the region. *LEARN Journal: Language Education and Acquisition Research Network Journal* 12(2): 28–48. *https://files.eric.ed.gov/fulltext/EJ1225665.pdf* (Last accessed: 28 April 2022).

Laurian, Simona–Carlton J. Fitzgerald. 2013. Effects of using rubrics in a university academic level Romanian literature class. *Procedia – Social and Behavioral Sciences* 76: 431–440.

Martyniuk, Waldemar–José Noyons. 2007. *Executive Summary of Results of a Survey on the Use of the CEFR at National Level in the Council of Europe Member States.* Strasbourg, France: Council of Europe.

North, Brian. 2020a. The CEFR renewed: Inspiring the future of language education. *Italiano LinguaDue* 12(1): 549–560.

—— 2020b. Trolls, unicorns and the CEFR: Precision and professionalism in criticism of the CEFR. *CEFR Journal—Research and Practice* 2: 8–24. *https:// cefrjapan.net/publications/journal* (Last accessed: 28 April 2022).

Piccardo, Enrica. 2019. *TIRF Language Education in Review – The Common European Framework of Reference (CEFR) in Language Education: Past, Present, and Future.* Monterey, CA & Baltimore, MD: TIRF & Laureate International Universities.

https://www.tirfonline.org/wp-content/uploads/2020/05/LEiR_CEFR.pdf (Last accessed: 28 April 2022).

Piccardo, Enrica–Brian North–Tom Goodier. 2019. Broadening the scope of language education: Mediation, plurilingualism, and collaborative learning: The CEFR companion volume. *Journal of e-Learning and Knowledge Society* 15(1): 17–36. *https://www.je-lks.org/ojs/index.php/Je-LKS_EN/article/view/1612* (Last accessed: 28 April 2022).

Schreiber, Lisa. M.–Gregory D. Paul–Lisa R. Shibley. 2012. The development and test of the public speaking competence rubric. *Communication Education* 61(3): 205–233. *https://www.tandfonline.com/doi/full/10.1080/03634523.2012.670709* (Last accessed: 28 April 2022).

Sundeen, Todd H. 2014. Instructional rubrics: Effects of presentation options on writing quality. *Assessing Writing* 21: 74–88. *https://www.sciencedirect.com/science/article/pii/S1075293514000191* (Last accessed: 28 April 2022).

Trinity College London. 2017. *Integrated Skills in English (ISE) Guide for Teachers – ISE II (B2) Reading & Writing | Speaking & Listening. https://www.trinitycollege.com/resource/?id=6292* (Last accessed: 29 April 2022).

Tuan, Luu T. 2012. Teaching and assessing speaking performance through analytic scoring approach. *Theory and Practice in Language Studies* 2(4): 673–679. *https://www.semanticscholar.org/paper/Teaching-and-Assessing-Speaking-Performance-through-Tuan/292d08af76f25e5cdf6b0a8a7e9c9187f2652e6d* (Last accessed: 28 April 2022).

Vez, José Manuel. 2011. La investigación en didáctica de las lenguas extranjeras. *Educatio Siglo XXI* 29(1): 81–108. *https://revistas.um.es/educatio/article/view/119881* (Last accessed: 28 April 2022).