Alicante Journal of English Studies / Revista Alicantina de Estudios Ingleses
ISSN: 0214-4808 | e-ISSN: 2171-861X
Issue No. 36, 2022, pages 85-109
https://doi.org/10.14198/raei.2022.36.05

Testing the Reliability of two Rubrics Used in Official English Certificates for the Assessment of Writing

Lucía Fraga Viñas

Author:
Lucía Fraga Viñas
Universidade da Coruña, Spain
lucia.fragav@udc.es
https://orcid.org/0000-0003-4602-3593

Date of reception: 20/01/2021
Date of acceptance: 06/07/2021

Citation:
Fraga Viñas, Lucía. 2022. "Testing the Reliability of two Rubrics Used in Official English Certificates for the Assessment of Writing." *Alicante Journal of English Studies* 36: 85-109.
https://doi.org/10.14198/raei.2022.36.05

Abstract:

The learning of English as a Foreign Language (EFL) is clearly a primary concern worldwide these days. This has spurred a proliferation of studies related to it and the emergence of new methodologies and instruments of assessment. Along with these, new qualifications devoted to the certification of language competence have been created, triggered in no small part by the fact that demonstrating one's level of proficiency has become almost an imperative when applying for a job or a grant, or to enable someone to study in a foreign country. It is therefore essential to test the reliability of the instruments used for the assessment of competences. With this purpose, over a four-week period, four different evaluators have assessed the written essays of students on a C1 level course using the writing rubrics for Cambridge Assessment English's Cambridge Advance English Certificate (CAE) and Trinity College's Integrated Skills in English Exams III (ISE-III). The aim was to examine the CAE and the ISE-III rubrics' reliability through the calculation of their respective Cronbach's alpha, the Corrected-Item Total correlation, the Intra-class Correlation Coefficient and the Standard Error of Measurement. Afterwards, the results given to each essay on the basis of the two rubrics were compared so to ascertain whether their language is clear and which criteria tended to obtain higher and lower marks on average. Examiners were also surveyed at the end of the assessment process to find their opinion on the use of the two rubrics in terms of clarity. The research provided meaningful and interesting results such as the fact that although both rubrics obtained good results in the coefficients of reliability, the variance in scores is greater

when using the ISE-III rubric and that examiners tend to be tougher when assessing the learner's language resource than any other criterion. It is also worth pointing out that according to the survey, examiners' general perception of both rubrics is that some of their descriptors were confusing or vague, which suggests both rubrics should be revised and could benefit from some improvement.

Keywords: rubrics; Official English Certificates; assessment; reliability.

1. Introduction

English is currently the main language used in trade and global communication and is therefore regarded as the world's *lingua franca*. As a result, it is also the most taught and studied language, and the teaching and assessment[2] of English as a Foreign Language[3] (hereafter, EFL) has become a matter of international importance.

With regard to the EFL classroom, traditional teaching methods have been gradually substituted by certain communicative methods. Since 2001, the European Council, through the *Common Framework of Reference for Language: Learning, teaching and assessment* (hereafter, CEFR), has also been promoting communicative competence and a communicative approach. This methodological shift triggered the need to find new assessment tools as the traditional paper tests based on grammatical activities no longer worked for the evaluation of communicative competence. Hence, new instruments of assessment like rubrics have come to be key in the evaluation processes that stem from the communicative approach.

In spite of the fact that many countries like the USA have been using rubrics as a common assessment tool since the beginning of the 20th century, far too little attention has been paid to them in Spain until recently. Their current presence in textbooks and evaluations has been prompted by the CEFR and the

2 Assessment and evaluation will be treated as synonyms in this article for stylistic convenience, although they are not exactly the same concept.

3 A foreign language, according to Richards and Schmidt, is a language which is not the native language of large numbers of people in a particular country or region, it is not used as a medium of instruction, and is not widely used as a medium of communication in government, media, etc. Foreign languages are typically taught as school subjects for the purpose of communicating with foreigners or for reading printed materials in that language (2002, 206). This is in contrast to a second language: a language that plays a major role in a particular country, though it may not be the first language of many people who use it (472).

education laws that have implemented it in our country. Research into rubrics or grading scales is still scarce in Spain, with many questions about its application, effectivity, and reliability yet to be addressed. Unfortunately, this leads to the common use of grading scales that are not valid, accurate or reliable to assess students, and thus they threaten the whole evaluation process.

Within the academic context, there are many institutions, both public and private, that deal with the granting of qualifications which certify an individual's level of proficiency in English. However, the determination of one's linguistic competence is highly complex and ardours, which complicates the evaluation process, already difficult per se, even further. Official English Certificates often use rubrics to assess writing and speaking skills. This is due to the fact that grading scales allows the examiner to measure, at the same time, different important aspects in a produced text with objectivity and precision. This is why it is so fundamental that the rubrics used are tested with the aim of finding out whether they are truly effective and reliable.

It is in this particular line of research where the current study fits as it attempts to test the reliability of rubrics used by two of the main providers of Official English Certificates: Cambridge Assessment English and Trinity College. In order to do so, four examiners have assessed various written texts using the grading scales of these institutions.

2. EFL Assessment

2.1. Historical Review

Assessment is not a new concept. In fact, according to Lavigne and Good; "forms of testing can be traced back to the [ancient] Chinese, Greek and Romans" (2014, 2). Nevertheless, it was not until the Middle Ages when examinations started to be much more formal. The 18th century saw an increasing demand to access education which would lead to an increase in evaluation, but only in form of entrance tests (Lavigne and Good 2014, 1-2). The education and evaluation processes at that time were completely different from how we currently conceive them, and were even far from our traditional concepts. Indeed, what we currently know as traditional schooling and testing was only born and developed in the 19th century. However, in that century only memory ability was tested.

Concerning EFL, Liz Hamp-Lyons (2016) locates the origins of formal large-scale examinations in the US to the period just before and just after the First World War. As for Britain, foreign languages were assessed with achievement purpose, just as ancient Greek and Latin had been examined. In 1911, the Cambridge University Senate suggested the creation of a teaching certificate

in modern foreign languages. In 1913, the Certificate of Proficiency in English was developed, prompted by an interest to improve Britain's relationships with colonies and former colonies. The test consisted of grammar and translation exercises as well as phonetic transcriptions, essays and pronunciation (Hamp-Lyons, 2016 15-17).

Another development in language testing and assessment took place in the late 1950s and early 1960s. John Carrol developed the Foreign Language Aptitude Battery which was designed to determine to what extent a person would be able to master a language. At around the same time, two proficiency tests were also developed in the United States: the certificate of Proficiency in English at the University of Michigan and the Proficiency Test of the American University Language Centre in Washington D. C, which would later develop into the now famous Test of English as a Foreign Language (TOELF). The American proficiency examinations were, however, different from the British one mentioned above, since they were influenced by advances in psychometrics and made no assumption of the learner having any previous knowledge of the language. In the same period, significant changes occurred in Britain, too. English was a very strong language owing to its importance in commerce and politics, and universities received thousands of international applications. It was therefore necessary to determine if a foreign student would be able to study in English. The English Proficiency Test Battery and the Test in English-Overseas were the two main examinations developed at this time (Hamp-Lyons 2016, 15-16).

The next change occurred in 1979 and was brought about following the appearance of Communicative Language Teaching. The British Council required a more communicative test to check proficiency within specific academic contexts. The English Language Testing Service (ELTS) was created, but it was too expensive, as well as hard to score and to carry out. It would be replaced by the International English Language Testing System (IELTS), which was more generic as it did not assess each individual according to the field he or she intended to study (Hamp-Lyons 2016, 17).

The 21st century was marked by the establishment of the CEFR (2001), which aimed to "overcome the barriers to communication among professionals working in the field of modern languages arising from the different education systems in Europe" (Council of Europe, cited in Hamp-Lyons 2016, 18).

In Spain, the introduction of foreign languages in school dates back only to the 20th century, when different reforms and education plans made the teaching of foreign languages compulsory for a few years at the time. The final major period started when the current Constitution was drafted (1978) and a new law was enacted (LOECE) two years later. It established the study of a foreign language, French or English, for all students in primary education. Since then, the political

parties in power have made changes in the education system through different laws: among others, the LOGSE, LOE and LOMCE (Morales et al. 2000).

2.2. Rubrics

Rubrics can be defined and understood in slightly different ways. Melissa D. Henning (2020, n.p.) defines a rubric as "a set of scoring guidelines that evaluate students' work and provide a clear teaching directive." According to Berkley University Center for Teaching and Learning (2020, para. 2), rubrics have four characteristics: the criteria students must achieve for a task, the indicator of quality which students should know and follow in order to pass the task (e.g., exceed expectation, meets expectation, doesn't meet expectation), the components or dimensions and the scale used as a scoring tool. However, the use of a rubric is also highlighted as beneficial, not only as a summative and formative tool, but also as a teaching tool which benefits teachers, students and the entire teaching-learning process. As Henning (2020, para. 2) explains:

> [Rubrics] "convey the teacher's expectations and they provide students with a concrete print out or electronic file showing what they need to do for the specific project. Typically, a teacher provides the rubric to students before an assignment begins, so students can use the rubric as a working guide to success"

Furthermore, they help the teacher during the assessment as they provide them with a complete range of criteria and goals in different aspects, not just grammatical. They also contain curriculum goals and standards.

Gavin Brooks (2012, 229) argues that rubrics were introduced in the L1 classroom in order to assess students' writing. Until that moment, writing tasks had been graded based on the criteria of the individual teacher, without any specific guidelines to support his or her decision. From the 70's on, rubrics also started to be used to give feedback (Brooks, 2012, 230).

Positive outcomes with regard to rubrics have been found in several studies. For instance, a study carried out in various educational centres in Spain concluded that, after using a rubric for a whole term in one subject, students considered that their motivation had increased and it boosted cooperative work (Gallego Arrufat & Raposo-Rivas, 2014).

Jonsson and Svingby (2007) did not conducted new research but instead revised seventy-five scientific studies on the reliability and validity of rubrics and concluded that grading is more consistent when they are used, and that both the reliability and validity of the assessment process increased. Similarly, Panadero and Jonsson (2013) analysed twenty-one studies, finding that rubrics provide

transparency to the assessment, reduce anxiety, aid with feedback and help to improve students' self-efficacy and self-regulation.

2.3. State-of-the art

Several studies have been carried out related to using rubrics in the assessment of writing skills, particularly in terms of the effect that showing learners the rubric in advance has. It is worth highlighting the research conducted by Todd Sundeen (2014) and Anthony Becker (2016). They both worked with several groups where some of the groups were shown the rubric and had it explained, while control groups were not. The results demonstrated the benefits on results of having access to the rubric in advance of tests. Another study conducted by Laurian and Fitzgerald (2013) also found that students obtained higher scores if they were shown the rubric previously.

In terms of research on the reliability of rubrics, Enayat A. Shabani and Jaleh Panahi (2020) examined the reliability of writing rubrics used by Official English Certificates [IELTS, Cambridge CAE, TOELF and Educational Testing Service (ETS] with a sample of 200 essays and four raters. It concluded there was very high agreement between test ratings and between raters. Moreover, the Cronbach's alpha calculated suggested high reliability, although some discrepancies were found between raters with respect to the Intra-class Correlation Coefficient (ICC). These results were in line with similar research that assessed the reliability of some rubrics used by official institutions through different coefficients (Fleckenstein et al. 2018; Trace et al. 2016; and Rupp et al. 2019).

Even though rubrics are used all over the world, their employment as an assessment tool in Spain is relatively recent. Research carried out by Velasco Martinez and Tojar in 2015 at Spanish universities indicated that from all the rubrics analysed, only 4% of the rubrics were used in the branch of arts and humanities, and they were used primarily for the assessment of essay writing (36%).

3. Methodology

3.1. Overview

In current society, a society of information and knowledge in an increasingly globalised world, students are required to obtain official qualifications that justify their knowledge of a language in an objective way. As Sundeen (2014, 79) explains, in this new "era of accountability," learners are subjected to several assessment processes at school, at the national level and even at the international level. It is in these last two scenarios where standardized tests come into play and

most students will be required to take them. For this reason, it is essential and fair that those standard assessment processes and their instruments of assessment are held accountable in terms of their reliability, validity and effectiveness.

3.2. Aims

This quantitative research aimed to calculate the Cronbach's alpha coefficient, the Corrected-Item Total correlation, the Intra-Class Correlation Coefficient (ICC) and the Standard Error of Measurement (SEM) of two of the writing rubrics used to grade two Official English Certificates at level C1: Cambridge Assessment English's Advanced Certificate (hereafter, CAE or Rubric 1) and Trinity College's Integrated Skills in English Exams (hereafter, ISE-III or Rubric 2). Furthermore, the scores obtained by each essay when scored by the same examiner (with Rubric 1 and with Rubric 2) will be compared. For instance, if one examiner gave the same essay a "5" using Rubric 1, and a "6" using Rubric 2. Likewise, a comparison among the scores each of the essays obtained with Rubric 1 and with Rubric 2 will also be made. For example, if an essay tended to get higher scores when being scored by one rubric. This will allow the variance between scores to be determined in order to check if both rubrics are assessing the same level of proficiency and how precise they are.

The comparison of scores by criteria (e.g. organisation, language, content) will also enable us to determine which criterion tend to be scored highest or lowest, on average. And finally, the survey conducted once the assessment process is over will allow conclusions to be drawn about the examiners' perceptions of both rubrics to examine whether they consider them to be clear or confusing, easy or difficulty to use, precise or not.

3.3. Participants

Four EFL examiners who are used to score with rubrics agreed to participate in the study and carried out a total of thirty-two assessments of writing using the two selected rubrics. The EFL examiners did not meet any of the participating students, whose names were removed from the writings for data protection purposes. The assessed writings were produced by students (ages ranged from 22 to 56 years old) from a C1-level course of instruction at the languages centre of the UNED University during the academic year 2019-2020 in A Coruña (Spain). Participants attended lessons two hours a week and had previously passed either the B2 course at the same languages centre or a B2 official certificate that permitted them to enrol in the C1 course, i.e., they had a level of proficiency beyond B2.

3.4. Ethical Procedures

The research conducted was carried out following the five ethical research procedures. Consent from all the research participants was obtained, the risk of harm to participants was minimized, anonymity and confidentiality of all participants was protected, no deceptive practices were employed and participants were given the right to withdraw from the research.

3.5. Procedures

The study was conducted over four weeks. At the outset, examiners were given clear instructions on what the study was to consist of and what they would need to do. They were also shown the rubrics and were explained to them what each of the criteria included in the rubrics measured. In the first week, examiners were sent their first two texts to grade, one with Rubric 1 and the other with Rubric 2. They were asked to complete an assessment chart indicating the rubric they were using, the text code used to identify the essay assigned, and the scores they considered appropriate for the text on the basis of each of the criteria in the rubric. The process was the same for the three following weeks. At the end, all the examiners had examined all the writing texts selected for the research twice, once with each of the rubrics. The study was specifically designed so that they did not have to score the same text in two consecutive weeks. This way examiners could distance themselves a little bit from their previous marking of the same text. For instance, examiner 1 assessed his/her two assigned texts in week 1 with their corresponding assigned rubric. Let's imagine that examiner 1 assessed essay A with Rubric 1 in week 1. He/she would assess the same essay A but with Rubric 2 in week 3. As a result, there are at least two weeks in between the first and second assessment of the same text so that his/her scoring is not influenced by his/her previous scoring with the other rubric.

At the end of the study, examiners were asked to complete a survey with twenty-six statements about their view on each rubric in terms of precision (whether they think the rubric is precise or not), the wording of the descriptors, number of criteria, scales (e.g. from 1 to 5, from 1 to 4, from perfect to deficient), etc. and also their experiences assessing texts with each of them. Examiners answered each question on a Likert Scale from 1 to 5, where 1 was "I strongly disagree" and 5 was "I strongly agree." In addition, they had space to write down any comments, observations, or clarifications.

3.6. Instruments

The reliability of the rubrics can be measured in different ways through different formulas and coefficients. In the current study, the measures used to assess instrument reliability were: Cronbach's alpha, the Corrected-Item Total Correlation, the ICC and the SEM.

Firstly, reliability can be measured in terms of internal consistency. For this purpose, the Cronbach's alpha, or coefficient alpha, was used. The Cronbach's alpha is a tool to evaluate the internal consistency or reliability of a set of scale or test items. According to Chelsea Goforth (2015, n.p), "the reliability of any given measurement refers to the extent to which it is a consistent measure of a concept, and Cronbach's alpha is one way of measuring the strength of that consistency." This index takes values between 0 and 1 and normally 0.7 is considered the minimum acceptable value. More specifically, a coefficient alpha above 9 is regarded as excellent, above 8 as good, above 7 acceptable, below 7 questionable, above 5 poor and below five unacceptable (George and Mallery 2003, 231). Concerning the minimum acceptable value, it must be clarified that most researchers consider 0.7 as the minimum acceptable value, although some authors such as van Griethuijsen et al. (2015) and Taber (2017) consider a coefficient of between 0.7 and 0.6 as acceptable; and for Goforth (2015, n.p), 0.65 is the minimum required. Nevertheless, it is also important to mention that the greater number of items a scale has, the more reliable the resulting coefficient will be.

Secondly, the Corrected-Item Total correlation indicates the corrected homogeneity coefficient. According to Faleye Bamidele Abiodun (2008, 833), this coefficient "indicates the new coefficient of 'Cronbach's Alpha' after a weak item had been removed from the scale. The set of items having low 'Corrected Item-Total Correlation' (of less than 0.2) are those that will increase the Cronbach's alpha coefficient of the scale when they are deleted".

Hence, the calculation of the Corrected-Item Total correlation allows the researchers to discover if the results of one specific item are interfering negatively in the reliability of the instrument as indicated by the Cronbach's alpha coefficient. If so, this item can be corrected, and the reliability of X (e.g. organisation, language) would improve. This coefficient has to be positive and above zero, otherwise it would mean that the item should be delated (Ciudad-Gómez & Valverde-Berrocoso, 2014).

The reliability of a rubric can also be checked through the analysis of the scores given by the different examiners with the same instrument. The ICC "measures the reliability of ratings or measurements for clusters — data that has been collected as groups or sorted into groups." (Glen, 2016). As such, an ICC close to 1 indicates high similarity while an ICC close to zero indicates the

values are not similar at all. Terry K. Koo and Mae Y. Lin (2016) explain that "based on the 95% confidence interval of the ICC estimate, values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively."

Moreover, in relation to the results obtained by students using one or another instrument, in this case, the rubric, the SEM "provides an indication of the dispersion of the measurement errors when you are trying to estimate students' true scores from their observed test scores." (Brown, 1999, 21) The SEM, in other words, indicates how close a test taker's score is likely to be to their 'true score', to within some stated probability.

The survey was specifically designed for this research following the techniques suggested by Juan Antonio Gil Pascual (2011, chapter 5). First, it was established that the questions or items in the survey would be items of a subjective numerical scale. Hence the scale contains twenty-six items and a Likert scale from 1 to 5, 1 being the lowest and 5 the maximum. The items are related to the examiners' perception of each of the rubrics in terms of the number of criteria, scale used, the wording of the descriptors, ease of use and time involved. Some "control" items were also included which made the same points but phrased differently with the objective of determining veracity. The first items in the survey were more general and easier to answer and items became more specific as the survey progressed. The coding of the survey was a table in which each of the items was divided by a line and alternative different colours (green and white) to facilitate the clarity of the table and to avoid participant errors when writing an X under the corresponding number of the Likert scale for each statement.

As far as the essays are concerned, all the essays assessed consisted in texts of between 240-260 words. The tasks, which always consisted of a prompt that students needed to develop, were included at the top of each essay. Students were given three ideas related to the prompt of which they should develop two. A proper introduction and conclusion were required.

> Example of one of the tasks:
> Which facilities should receive money from local authorities?
> - Public gardens
> - Museums
> - Sport Centres
>
> Write an essay discussing **two** of the facilities. You should explain which facilities you think are more important for the local authorities to consider, giving reasons to support your opinion.

Finally, both rubrics employed are analytic and not holistic. This is presumably due to the fact that analytic rubrics allow different aspects of the text that has been produced by the candidate to be measured, for instance, organization, lexical resource, grammatical resource, coherence and cohesion, etc. The results obtained are therefore much more precise than those obtained with a holistic rubric, as research has shown (Sundeen 2014; Becker 2016). Rubrics can be classified according to its application, in that case the Cambridge's rubric is skill-focused while the Trinity's one is a task-focused one. The reason why is that in the CAE the same rubric is used to score all the writing tasks while in the ISE-III each of the writing tasks of the exam use a different rubric. The rubrics have different criteria in terms of number, though they are similar in content, both following CEFR indicators of level for C1 writing. While the writing rubric of the ISE-III consists of three criteria: "task fulfilment", "organisation and structure" and "language control", the CAE writing rubric has four: "content", "communicative achievement", "organisation" and "language". However, the "task fulfilment" criterion of the ISE-III encompasses the same aspects as the CAE criteria of "content" and "communicative achievement" measure.

3.7. Data analysis

The in-depth analysis of the data was carried out with the IBM statistics software SPSS, which calculates, together with the above cited Cronbach's alpha coefficient, Corrected-item Total Correlation and the ICC, and other coefficients and formulas that provide the author with in-depth descriptive analysis of frequency such as standard deviation, range, variance, etc.

The formula of the Cronbach's alpha coefficient is shown below, N is equal to the number of items, c is the average inter-item covariance among the items and v equals the average variance:

$$\text{~} \quad \frac{N\bar{c}}{\bar{v} + (N-1)\bar{c}}$$

The formula for the ICC is:

$$ICC = \frac{S^2_b}{(S^2_b + S^2_w)}$$

where S^2w is the variance within subjects, and S^2b is the variance of measurements between subjects. S^2b+S^2w is the total variances. As such, the ICC is interpreted as the proportion of total variance accounted for by the within subject variation. (Howell 2018, n.p)

SEM is calculated with the following formula: $SEM = S\sqrt{1-r_{xx}}$ where S stands for Standard Deviation and rxx stands for the coefficient of reliability, in this case, Cronbach's alpha.

In order to make the comparison of the scores obtained with each of the rubrics and the scores obtained by one essay when being assessed with Rubric 1 or Rubric 2 by each of the examiners, the variance was measured, and the means were obtained. Sample variance is obtained with the following formula $Sample\ variance = \frac{\sum(x-\bar{x})^2}{(n-1)}$ and the mean with $Mean\ \bar{x} = \frac{\sum xi}{N}$, although IBM's statistics software SPSS and Microsoft Excel software can calculate them automatically once the data are introduced.

3.8. Hypotheses

The hypotheses of this study are:

H1.-The CAE rubric will obtain an excellent Coefficient Cronbach's alpha.

H2.-The ISE-III will not obtain an excellent Coefficient Cronbach's alpha.

H3.-The Corrected-Item Total Correlation will indicate heterogeneity in both rubrics.

H4.-Both rubrics will obtain a deficient Intra-class Correlation Coefficient.

H5.-The criterion related to language grammar will obtain the lowest scores.

4. Findings

The rigorous analysis on the data obtained from the examiners' scoring of the texts together with the survey conducted afterwards revealed some interesting results. These are presented below following the order of the research hypotheses.

As has already been explained, the data were examined in multiple ways in order to determine the Cronbach's alpha of each of the rubrics, the variance among scores, which criteria tend to obtain the higher or the lower marks, in which criteria there were bigger discrepancies between the raters, etc.

To begin with, the Cronbach's alpha of each rubric was calculated. Each of the texts was assessed by each of the examiners with both rubrics.

Table 1. Cronbach's alpha for Rubric 1

Cronbach's alpha	Cronbach's alpha based on standardized elements	No of criteria used in the rubric (e.g. content, language, organization)
,951	,963	4

As it can be observed in the table above, the Cronbach's alpha of Rubric 1 shows high reliability because it is 0.951, so it is regarded as excellent.

Table 2. Cronbach's alpha for Rubric 2

Cronbach's alpha	Cronbach's alpha based on standardized elements	No of criteria used in the rubric
,928	,938	3

The Cronbach's alpha for Rubric 2 also showed high reliability with a value 0.928 although it is slightly lower than the value for Rubric 1 (0.951), so it is not statistically significant. These two analyses prove high internal consistency in both rubrics. Consequently, H1 is supported by the results while H2 is not.

In terms of the Corrected-Item Total Correlation, if it is zero or negative for an item, the item should be deleted.

Table 3. Analysis of Corrected-Item Total Correlation in Rubric 1

	Scale mean if the item is deleted	Scale variance if the item is deleted	Total Corrected item correlation	Cronbach's alpha if the item is deleted
Content	8,9375	13,796	,902	,948
Communicative Achievement	9,1250	18,917	,861	,947
Organisation	9,0000	16,000	,933	,920
Language	9,3125	18,629	,939	,930

As the table shows, all the Cronbach's alpha values if one item is deleted are similar, which implies high internal consistency. The Corrected-Item Correlation values are not zero or negative so none of the items should be deleted.

Table 4. Analysis of Corrected-Item Total Correlation in Rubric 2

	Scale Mean if the item is delated	Scale variance if the item is delated	Total Corrected Item Correlation	Cronbach's alpha if the item is delated
Task Fulfilment	5,1250	3,983	,886	,883
Organisation and Structure	5,0625	4,329	,920	,839
Language Control	5,3125	5,962	,821	,945

With regard to Rubric 2, all the Cronbach's alpha values if one item is delated show high reliability as they are all above 0.8 and they are similar too. The Corrected-Item Correlation values are far from being zero or negative so none of the items should be delated in this case either. The results concerning Corrected-Item Total Correlation of both rubrics refute H3.

Another objective of the study was to check the scores obtained by the same text with the two rubrics. Concerning the concordance between the examiners' scores of the same text using Rubric 1, the results show variance of between 0.10 and 1.17 while with Rubric 2, it is between 1.6 and 2.6. It should also be highlighted that on three occasions, the same text obtained the same score with Rubric 1 and Rubric 2 when being assessed by the same examiner. That said, examiners tended to give a higher mark to essays when using Rubric 2 (on 8 out of 13 occasions).

All the texts were assessed by all the examiners over the four weeks of the study, first with one rubric and then with the other. The scores given by the four examiners for each essay using the same rubric were analysed so that ICC could be calculated.

Table 5. ICC of text scores assessed with Rubric 1.

ICC

	Intra-class Correlation	**95% confidence interval**	
		Lower Limit	Upper Limit
Unique measurements	,735	,303	,977
Mean measurements	,917	,635	,994

The ICC value for Rubric 1 indicates that there is excellent reliability since the coefficient is above 0.9. This supports the reliability of the instrument since a very unreliable rubric would produce high discrepancy in the examiners' scores.

Table. 6. ICC of text scores assessed with Rubric 2.

	ICC		
		95% Confidence Interval	
	Intra-class correlation	Lower Limit	Upper Limit
Unique measurements	,652	,198	,967
Mean measurements	,882	,497	,992

The ICC value for Rubric 2 indicates that the intra-class correlation is good since the result is above 0.8. H4 is not supported by the findings of the research.

The SEM obtained with the results from Rubric 1 is 0.722 and for Rubric 2 is 0.730 meaning that the scores obtained are reasonably close to the student's true scores in terms of probability.

The study also considered the results of each rubric by criteria so that those criteria which tended to receive higher or lower scores could be found, as well as which level of the scale was selected more frequently for each criterion.

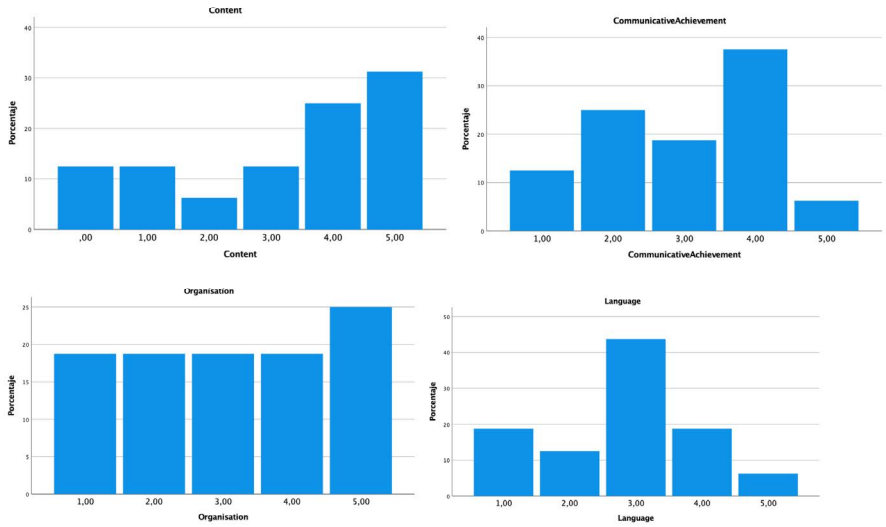
Beginning with Rubric 1, the criterion "content" was on average scored the highest while "language" was the lowest, as can be seen in Table 7.

Table 7. SPSS statistics of Rubric 1 scores for each criterion.

		Content	Communicative Achievement	Organisation	Language
N	Valid	16	16	16	16
	Missed	0	0	0	0
Mean		3,1875	3,0000	3,1250	2,8125
Standard Error of the mean		,45843	,30277	,37500	,29182
Desviation		1,83371	1,21106	1,50000	1,16726
Variance		3,363	1,467	2,250	1,363

A descriptive analysis of the frequency shows that the criterion "Content" was scored with a 5 or a 4 in 56% of the assessments and with a 1 in only 6.3%. The criterion "Communicate Achievement" was scored with a 4 on 37% of occasions and with a 5 in just 6.3%. Regarding "Organisation", a 4 was given in 25% of the assessments. Finally, "Language" was scored with a 3 in 43% of cases and with a 5 only in 6.3%.

FIGURE 1. Analysis of frequency of different scores using Rubric 1.



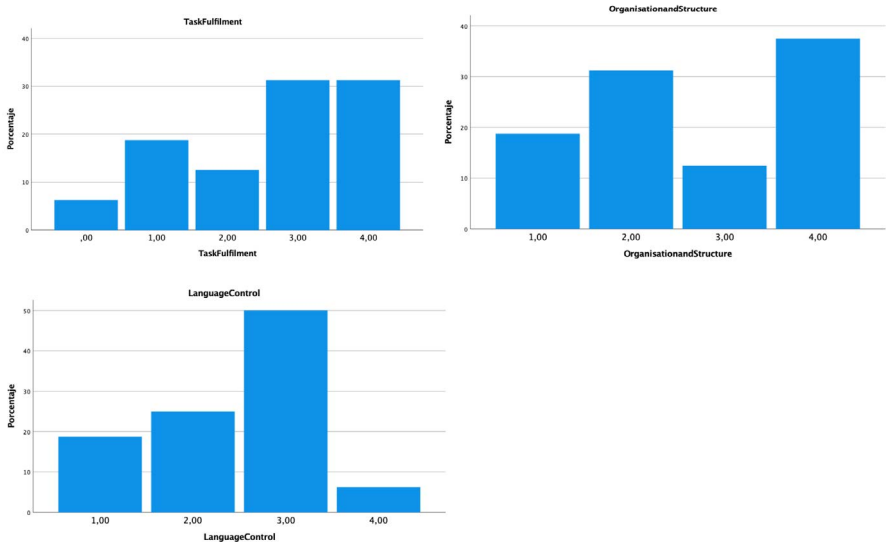
As far as Rubric 2 is concerned, the criterion "Organisation and Structure" was scored the highest on average while "Language Control" obtained the lowest marks. The data obtained related to "language control" in both rubrics support H5.

FIGURE 2. Analysis of Rubric 2 scores by criteria

		Task Fulfilment	Organisation and Structure	Language Control
N	Valid	16	16	16
	Missed	0	0	0
Mean		2,6250	2,6875	2,4375
Standard error of the mean		,32755	,29887	,22302
Desv.		1,31022	1,19548	,89209
Variance		1,717	1,429	,796

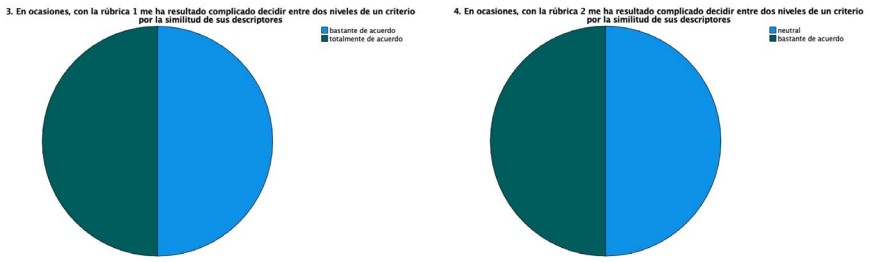
In terms of frequency, "Task Fulfilment" received either a "3" or a "4" in 62.6% of the assessments and a "0" in 6.3%. The criterion "Organisation and Structure" was scored with a "4" in 37.5% of the times and with a "3" in only 12.5%. Finally, "Language Control" obtained a "3" in 50% of the cases and a "4" in 6.3%.

FIGURE 3. Analysis of frequency of each score by criterion using R2.



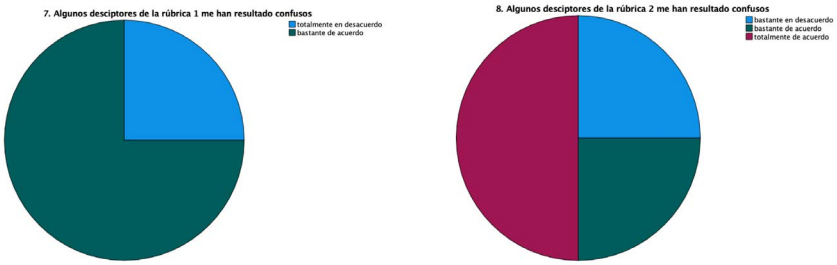
With respect to the survey conducted after the assessments, all the examiners expressed strong agreement or agreement with statements 3: "sometimes it has been difficult for me to decide the score between two levels for one criterion because the descriptors were too similar" with Rubric 1 while only 50% agreed with the same statement for rubric 2 (st. 4).

FIGURE 4. Survey results for statements 3 and 4



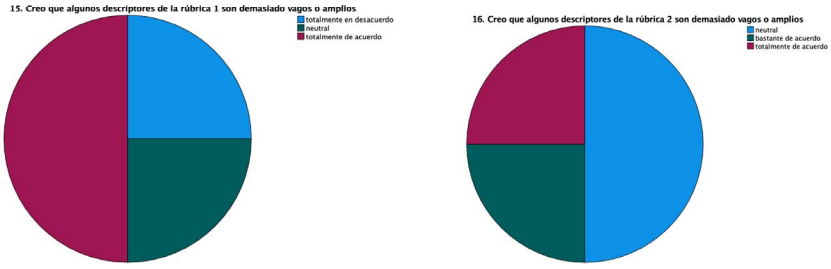
Moreover, 75% of the examiners agreed that the descriptors of Rubric 1 were confusing (st. 7) and 75% also agreed or strongly agreed that the descriptors of Rubric 2 (st. 8) were confusing.

FIGURE 5. Survey results for staments 7 and 8



Concerning vagueness, 50% of the examiners strongly agreed that the descriptors in Rubric 1 (st. 15) were vague, the same percentage agreeing or strongly agreeing with the same statement for Rubric 2 (st. 16).

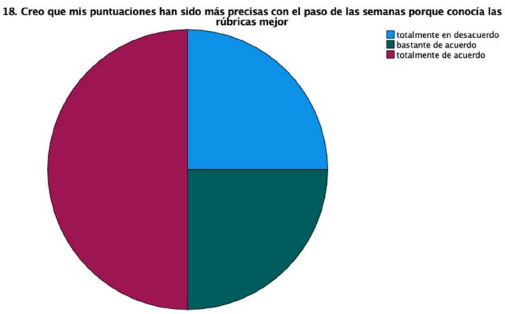
FIGURE 6. Survey results for statements 15 and 16.



Other interesting findings obtained were the fact that 50% of the examiners wished the scale of Rubric 2 (from 0 to 4) was bigger (st. 6), and 75% that the same rubric assessed more than just three criteria (st. 12).

And finally, 75% of the examiners either agreed or strongly agreed that their assessments became more accurate over time as they felt they knew the rubrics better (st. 18).

FIGURE 7. Survey results for statement 18



5. Conclusions

The research conducted allows reflections to be made related to many different aspects. The Cronbach's alpha coefficients calculated for the two rubrics show high indices of reliability as both rubrics scored over 0.9, which is normally considered excellent. However, as has already been noted, Cronbach's alpha index is more reliable when there are more items. In the cases analysed, the rubrics consist of either three or four items. This is very significant, particularly for the Trinity College rubric since it only measures three criteria, as it could lead to a falsely high value of coefficient alpha.

The Corrected-Item Total Correlation for the items in the CAE and ISE-III rubrics suggests internal consistency among the items of each of the two rubrics, which is slightly higher in the case of the former, where all the alpha values were above 0.9.

Some rather revealing conclusions can be drawn from the analysis of scores. First of all, the ICC for both the CAE and the ISE-III rubric were high, although it was a little better for the former. This indicates that examiners' scores using the same rubric were similar, which indicates high reliability. By comparing the scores given by the four examiners text by text, it was found that the average difference in scores awarded using the CAE rubric oscillated between 0.10 and 1.17 points, whereas with the ISE-III rubric, scores varied by 1.6 to 2.6 points. This fact suggests that scores with Rubric 1 are more precise. The variance in the ISE III rubric suggests a text could be either failed or passed with the same rubric, perhaps indicating a certain lack of exactness.

The contrast between the scores given by the same examiner to the same text with each of the rubrics indicates that examiners tend to mark a text higher when they are assessing it with the Trinity rubric than with the Cambridge

one. However, the variance is less than 0.5, which suggests that they are truly assessing the same level of proficiency.

Considering the raters' scores by criteria provides some interesting data and allows some reflections. In this respect it is worth highlighting that the criterion "Language" was always scored the lowest on average with both rubrics. In the total of thirty-two assessments that were carried out for this research, it was only given the highest score of "5" for Rubric 1 and "4" for Rubric 2 on four occasions. This raises the question of whether examiners are stricter when assessing the grammatical and lexical range of the learner than when they are evaluating other criteria. In contrast, examiners gave high scores in the criterion "Content" (Trinity) or "Task fulfilment" (Cambridge) in eighteen out of thirty-two assessments. Therefore, it appears that students are much more likely to obtain a high mark in the "Content/Task fulfilment" than in the "Language" criterion.

As far as the survey on the examiner's opinion upon the rubrics is concerned, some of their perception's should be mentioned. With regard to the Cambridge CAE rubric, all raters stated that it had been difficult for them to decide between two levels of one criterion because the descriptors were sometimes too similar. There is no doubt that in order for a rubric to be good and effective, the different levels of the scale should be clearly differentiated. This would increase the rubric's reliability and the ICC. In addition, 75% of the examiners believe the descriptors of Rubric 1 are confusing and half agree that they are too vague. In fact, some examiners have indicated in the comments section that they dislike the fact that levels 2 and 4 are not really worded since it is just said they correspond to an intermediate level between the band above and below.

The results of the survey regarding the Trinity ISE-III rubric show that half of the examiners wished the scale were bigger since it goes only from 0 to 4. They also think that the descriptors of the rubric are confusing and are often too similar for the different levels of the same criterion, although to a lesser extent than for Rubric 1. Moreover, it must also be emphasized that examiners do not regard the ISE-III descriptors as vague. In spite of this, half of them considered that it would consume too much time to use this rubric with a large number of students and 75% think this rubric is rather imprecise.

5.1. Comparison with the Findings of Other Studies

The results of this research are in line with those obtained by the study conducted by Shabani and Panahi (2020) where the official writing rubrics analysed, which included the CAE writing rubric, obtained high coefficients of

reliability. However, some variance in the ICC was found particularly with two of the raters. In that study, the ISE-III rubric was not analysed.

Moreover, the conclusions drawn from the results are similar to another research on the same topic. Similar to Trace et al. (2016), it was found that raters often show discrepancies in their understanding of descriptors. Rupp et al.'s (2019) research on the reliability of human raters also obtained high reliability coefficients among writers when using TOEFL rubrics.

Research provided by Cambridge English Assessment (2020) indicates that their measurements of Cronbach's alpha of examiners marking the CAE writing paper is 0.79 and its SEM 1.78. In this case, the results obtained by the current research concerning Cronbach's alpha are even higher (above 0.9). As for the SEM data, the results of the current study show even lower results, around 0.7, which is very positive.

No data is published on Trinity College webpage regarding reliability coefficients. Moreover, no research on the reliability of ISE-III was found in the literature, so it is not possible to compare with the findings of this research.

5.2. Research Limitations

Despite the fact that the research conducted has led to many interesting findings, they must be taken with caution, since there are multiple limitations to this study. The main limitation of the research carried out is its size. It is clear that the larger the sample of essays is, the more reliable, accurate and significant the results obtained will be. It would be strongly recommendable to reproduce the same research on a larger scale by assessing writing samples of, for instance, the participants of C1 English instruction courses of different universities.

On the other hand, the rubrics selected are supposed to be based on exhaustive investigations by an entire panel of experts, so they must be taken with the presupposed reliability they deserve. It is essential, though, to consider that they have been developed years ago and thus they may be subject of adjustments, improvements and modifications.

5.3. Implications and Future Research

Since 75% of the examiners think that their scoring became more precise as time passed because they knew the rubrics better, the relevance of training raters on the assessment with one rubric is highlighted. Even though it is presupposed that examiners for Cambridge and Trinity Certificates are indeed trained for their assessment, there are other teachers who use rubrics in their

courses or education centres that are not properly trained to assess with them. As a result, if they received some training in their university degrees or masters, examiner's own accuracy in their scoring could be raised, which will be obviously extremely beneficial for the whole evaluation process. Therefore, further research on the benefits of examiners training with rubrics is desirable, together with the analysis of the impact that a proper training of the examiners can actually have.

With regard to the findings of the research, the rubrics assessed could benefit from some adjustments. As suggested by examiners, some of the descriptors of both rubrics could be rephrased as they are often found confusing. Raters suggested that the CAE rubric is too vague as some of the bands are not really described and this is something that could be corrected. Regarding the ISE-III rubric, the examiners have emphasized the fact that the descriptors are confusing, this could obviously result in decreased reliability, so the descriptors should be reviewed. In addition, the descriptors of both rubrics have been described as too similar between the different scale levels, which could also be addressed in future revisions of the rubrics. Further research on the wording of descriptors could be conducted so that some light could be shed on how to design better rubrics that are clear, accurate and precise.

Due to the limitations of the current research that have been already mentioned, further research on the reliability of rubrics used in official English certificates with a larger number of participants is strongly recommended and could be very beneficial for the educative community and the assessment process. After all, in this era of accountability it should be compulsory that students are not only the object of a fair evaluation to demonstrate their knowledge, competence or ability, but also the processes of assessment and the instruments used for it should be subjected to analysis, examination and revision from time to time through research, so that it can be demonstrated that they are truly objective and effective.

Works Cited

Becker, Anthony. 2016. "Student-generated scoring rubrics: Examining their formative value for improving ESL students' writing performance." *Assessing Writing* 29: 15-24. https://doi.org/10.1016/j.asw.2016.05.002

Berkeley University Center for Teaching & Learning. 2020. "Rubrics." *teaching.berkeley.edu*. Accessed online on the15th of July 2020: https://teaching.berkeley.edu/resources/assessment-and-evaluation/design-assessment/rubrics

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* Vol. 17: 227-240. Accessed online on the 20th of July 2020: core.ac.uk/download/pdf/143638458.pdf

Brown, James D. 1999. "Questions and answers about language testing statistics: Standard error vs. Standard error of measurement." *Shiken: JALT Testing & Evaluation SIG Newsletter*, 3 (1): 20-25. Accessed online on the 5th of August 2020: http://hosted.jalt.org/test/bro_4.htm

Cambridge English Assessment. 2020. "Quality and accountability". *Cambridge English* webpage. Accessed online on the 4th of August 2020: https://www.cambridgeenglish.org/research-and-validation/quality-and-accountability/

Ciudad-Gomez, Adelaida and Jesús Valverde-Berrocoso. 2014. "Reliability Analysis of An Evaluation Rubric For University Accounting Students: A Learning Activity About Database Use." *Journal of International Education Research (JIER)* 10(5): 301-306. Faleye, Bamidele Abiodun. 2008. "Reliability and Factor Analyses of a Teacher Efficacy Scale for Nigerian Secondary School Teachers." *Electronic Journal of Research in Educational Psychology* 16, Vol 6 (3): 823 – 846. Fleckenstein, Johanna, Stephan Keller, Maleika Kruger, Richard J. Tannenbaum, and Olaf Köller. 2019. "Linking TOEFL iBT writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study." *Assessing Writing* 43. https://doi.org/10.1016/j.asw.2019.100420

Gallego Arrufat, María Jesús and Manuela Raposo-Rivas. 2014. "Compromiso del estudiante y percepción del proceso evaluador basado en rúbricas." *REDU. Revista de docencia universitaria*. 12, 1: 197-215. George, Darren and Paul Mallery. 1995. *SPSS/PC+step by step: A simple guide and reference*. United States: Wadsworth Publishing Company.

Gil Pascual, Juan Antonio. 2011. *Técnicas e Instrumentos para la recogida de información*. Madrid: Universidad Nacional de Educación a Distancia.

Glen, Stephanie. 2016. "Intraclass Correlation." *StatisticsHowTo.com: Elementary Statistics for the rest of us!* Accessed online on the 6th of August 2020: www.statisticshowto.com/intraclass-correlation/

Goforth, Chelsea. 2015. "Using and interpreting Cronbach's Alpha." *University of Virginia. Research Data Services + Sciences*. November 16. Accessed online: https://data.library.virginia.edu/using-and-interpreting-cronbachs-alpha/

Hamp-Lyons, Liz. 2016. "Purposes of Assessment." In Tsagari and Banerjee (ed(s).). *Handbook of second language assessment*, The Hague, De Gruyter/Mouton, pp. 13-28. Henning, Melissa. D. 2020. "Rubrics to the Rescue: What are rubrics?" *TeachersFirst. Thinking Teachers Teaching Thinkers*. The Source of Learning, Inc. Accessed online on the 6th of July 2020: www.teachersfirst.com/lessons/rubrics/what-are-rubrics.cfm

Howell, David. 2018. "Intraclass Correlation: Multiple Approaches" *University of Vermont*. Outline of the Statistical Pages Folder. Accessed online on the 28th of July 2020: https://www.uvm.edu/~statdhtx/StatPages/icc/icc-overall.html

Jonsson, Anders and Gunilla Svingby. 2007. "The use of scoring rubrics: Reliability, validity and educational consequences." *Educational Research Review* 2: 130–144. Koo, Terry. K and Mae Y. Lin. 2016. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine* 15, 2: 2016, 155-163. Lavigne, Alyson L. and Thomas L. Good. 2014. *The Teacher and student evaluation: moving beyond the failure of school system*. New York, Routledge.

Laurian, Simona and Carlton Fitzgerald. 2013. "Effects of using rubrics in a university academic level Romanian literature class". *Procedia. Social and Behavioral Sciences* 76: 431-440. https://doi.org/10.1016/j.sbspro.2013.04.141

Morales, Carmen, Laura Ocaña Villuendas, Alicia López Gayarre, Irene Arrimadas Gómez and Eulalia Ramirez Nueda. 2000. *La enseñanza de las lenguas extranjeras en España*. Secretaría General Técnica. Centro de Publicaciones. Ministerio de Educación, Cultura y Deporte. Accessed online on the 27th of June 2020: sede.educacion.gob.es/publiventa/la-ensenanza-de-las-lenguas-extranjeras-en-espana/investigacion-educativa/8757

Panadero, Ernesto and Anders Jonsson. 2013. "The use of scoring rubrics for formative assessment purposes revisited: A review." *Educational Research Review* 9: 129–144., Richards, Jack C. and Richard Schmidt. 2002. *Language Teaching & Applied Linguistics*, Longman, Pearson Education.

Rupp, André A., Jodi M. Casabianca, Maleika Krüger, Stephan Keller, and Olaf Köller. 2019. "Automated essay scoring at scale: a case study in Switzerland and Germany" (RR-86. ETS RR-19-12). *ETS Research Report Series*. Shabani, Enayat A. and Jaleh Panahi. 2020. "Examining consistency among different rubrics for assessing writing". *Language Testing in Asia* 10:12. https://doi.org/10.1186/s40468-020-00111-4

Sundeen, Todd. H. 2014. "Instructional rubrics: Effects of presentation on writing quality." *Assessing writing* 21: 74-87. https://doi.org/10.1016/j.asw.2014.03.003

Taber, Keith. 2017 "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education." *Research in Science Education* 48: 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Trace, Jonathan, Valerie Meier, and Gerriet Janseen. 2016. "I can see that: developing shared rubric category interpretations through score negotiation." *Assessing Writing* 30: 32–43 .

Tsagari, Dina and Jayanti Banerjee, eds. 2016. *Handbook of Second Language Assessment*. The Hague, Walter de Gruyter Inc.

Van Griethuijsen, Ralf, Michiel W. van Eijck, Helen Haste, Perry J. den Brok, Nigel C. Skinner, Nasser Mansour, Ayse Savran Gencer and Saouma BouJaoude. 2015. "Global Patterns in Students' Views of Science and Interest in Science." *Res Sci Educ* 45: 581–603. https://doi.org/10.1007/s11165-014-9438-6

Velasco-Martínez, Leticia and Juan Carlos Tójar. 2015. "Evaluación por competencias en educación superior. Uso y diseño de rúbricas por los docentes universitarios." AIDIPE (Ed.), *Investigar con y para la sociedad* 2: 1393-1405. Accessed online on the 17th of June 2020: avanza.uca.es/aidipe2015/libro/volumen2.pdf