The Use of the Common European Framework of Reference for Languages to Evaluate Compositions in the English Exam Section of the University Admission Examination

El uso del Marco Común Europeo de Referencia para las Lenguas para evaluar las redacciones en la sección de inglés de la Prueba de Acceso a la Universidad[1]

María Belén Díez-Bedmar

Universidad de Jaén. Facultad de Humanidades y Ciencias de la Educación. Departamento de Filología Inglesa. Jaén, España.

Abstract

The Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) has become the standard used to describe and evaluate students' command of a second or foreign language. However, it has not been used yet to evaluate students' command of English as a foreign language before entering university, i.e., when taking the English exam portion of the University Admission Examination. This paper aims at bridging this gap in the literature by means of a twofold objective. First, to use the CEFR to evaluate the compositions written for the University Admission Examination and explore the level of proficiency displayed there. Second, to analyse inter-rater reliability when raters use the CEFR, so as to unveil the problems raters find in actual use. A representative sample of the compositions written on one topic in the exam

[1] The author would like to acknowledge the help provided by the *Comisión Interuniversitaria Andaluza para las Pruebas de Acceso a la Universidad,* and specially to its Secretary Dr. Bengoa Díaz, for allowing her to have access to the English exams written for the University Entrance Examination in June 2008 in Jaén. Thanks are also due to the *Vicerrectorado de Estudiantes e Inserción Laboral at the Universidad de Jaén,* and Drs. Bueno González and Pérez Paredes. Similarly, the author would like to express her gratitude for making the publication of this paper possible to the project *«El sistema de acceso a la Universidad: propuestas en la gestión, decisión e inferencias en el área de lenguas extranjeras»* (FFI2011-22442), funded by the *Ministerio de Educación.*

Fecha de entrada: 10-07-2011 Fecha de aceptación: 25-10-2011

was selected, and two raters were asked to evaluate them according to the CEFR. The results in this article show that inter-rater reliability is very low (k= .245) when raters use the CEFR The findings also highlight that, in the cases where there was total inter-rater agreement, most of the compositions (91.33%) were placed at the B1 level. The results of this paper can inform raters, test designers and education authorities of the marks awarded to student compositions at this stage. Moreover, the findings highlight the rating aspects which would have to be revised or adapted if the CEFR is to be used in the future to mark compositions in the English exam portion of the University Admission Examination.

Keywords: evaluation, written expression, foreign languages, Common European Framework of Reference for Languages, proficiency levels, University Admission Examination, inter-rater reliability.

Resumen

El Marco Común Europeo de Referencia para las Lenguas (MCER) (Consejo de Europa, 2001) se ha convertido en el estándar para describir y evaluar el dominio que los estudiantes tienen en una segunda lengua o lengua extranjera. Sin embargo, no se ha utilizado aún para evaluar el dominio de inglés que los alumnos tienen antes de entrar en la universidad, es decir, cuando escriben su examen de Inglés en las Pruebas de Acceso a la Universidad. Este artículo tiene como objetivo cubrir esta laguna existente por medio de un doble objetivo. Primero, usar el MCER para evaluar las redacciones escritas en dicho examen y explorar los niveles en los que se encuentran las redacciones. En segundo lugar, analizar la fiabilidad interjueces cuando se utiliza el MCER para descubrir los problemas que se encuentran al utilizarlo. Para poder alcanzar estos objetivos, se seleccionó una muestra representativa de las redacciones escritas sobre un tema en el examen, y se pidió a dos evaluadoras que clasificaran las redacciones de acuerdo con el MCER. Los resultados de este artículo muestran que la fiabilidad interjueces es muy baja (k= ,245) cuando se utiliza el MCER. Los datos obtenidos también apuntan a que, teniendo en cuenta los casos en los que las dos evaluadoras tuvieron acuerdo total, la mayoría de las redacciones escritas por los alumnos (91,33%) se encuentra en el nivel B1. Estos resultados pueden informar a los evaluadores, los diseñadores de pruebas y a las autoridades educativas sobre los niveles otorgados a las redacciones. Los datos aportados también señalan qué aspectos necesitarían revisión o adaptación si el Marco Común de Referencia de las Lenguas se va a utilizar en el futuro para evaluar las redacciones en el examen de inglés de la Prueba de Acceso a la Universidad.

Palabras clave: evaluación, expresión escrita, lenguas extranjeras, Marco Común Europeo de Referencia para las Lenguas, niveles, Prueba de Acceso a la Universidad, fiabilidad interjueces.

Fecha de entrada: 10-07-2011 Fecha de aceptación: 25-10-2011

Introduction

The establishment of the European Higher Education Area (EHEA), which ensures the transparency and comparability of qualifications in Europe, the academic recognition of the studies taken in any European university (i.e. mobility programmes), and promotes the students' access to other undergraduate or graduate programmes abroad, demanded the establishment of a yardstick which would be used to determine the students' command of a foreign language (FL, henceforth).

The Common European Framework of Reference for Languages (CEFR, henceforth) (Council of Europe, 2001) was the framework chosen to describe the language standards which are to be attained by the students in their secondary or higher education. In the Spanish context, the CEFR levels are used to determine the levels to be achieved at a very important moment in a student's life: the University Entrance Examination (UEE, henceforth). However, a contradiction is found between the use of the CEFR levels and their use for the assessment of the language produced in the English exam in the UEE: whereas the CEFR levels are stated in the legal documents, the evaluation of the exam is not carried out by following the CEFR and the Can-Do statements. Therefore, the use of the CEFR scales and the Can-Do statements to evaluate the English exam may prove the following step in the possible changes to be implemented in this exam in the UEE.

Although there have been previous studies on the English exam in the UEE (for overviews, see García Laborda, 2006, and Díez-Bedmar, in press), and on the students' use of the FL in the English exam (Díez-Bedmar, in press), the evaluation of their written expression by means of the CEFR has not been explored yet. For this reason, this paper aims at bridging this gap in the literature by means of a twofold objective. First, to use the CEFR to explore the CEFR levels which are awarded to a representative sample of the compositions written in the English exam in the UEE. Thus, it will be possible to know if students meet the language expectations in the official documents. Second, to analyse the inter-raters' reliability when using the CEFR to unveil the problems found when raters working at secondary and higher education (as is the case nowadays) use the CEFR.

The results of this paper may be used to inform the designers of the English exam, their raters, and the education authorities of the CEFR levels that the students show at this stage, and of the rating aspects which will need revision or adaptation if the CEFR and the Can-Do Statements are to be used in the future to evaluate the English Exam in the UEE.

Fecha de entrada: 10-07-2011 Fecha de aceptación: 25-10-2011

Literature Review

Two main blocks will be developed here. The first one will offer a literature review of the CEFR levels which are set as requirements for students in Spain and abroad. The second block will be devoted to a review of the main sources of errors which are found when evaluating tests, regarding the type of rating scale and the raters.

The Common European Framework of Reference for Languages: Levels and Requirements

The CEFR is a positively worded analytic assessor-oriented, diagnosis-oriented or user-oriented scale (Alderson, 1991), which provides a grid with descriptors of communicative activities as well as descriptors of aspects of proficiency related to particular competences. These are presented in both vertical and horizontal dimensions. In the former, the raters grade whether the students meet the functional Can-Do criteria for each level, usually with the help of benchmark papers, which results in the allocation of the paper into a CEFR level. However, in the second, the raters check whether the students meet the specific criteria regarding aspects related to range, coherence, accuracy as well as specific Can-Do statements related to the text genre. For this reason, it can be said that analytic criteria scales are used, but the rating is holistic, since the rater needs to make a judgement by matching the criteria with the student's actual performance. Despite the advantages of the CEFR, many aspects of its design have been criticised. Among them, the need for further explicitation of the levels has been highlighted (Alderson et al., 2004; cited in Weir, 2005; Huhta et al, 2002; Kaftandjieva and Takala, 2002).

In the Spanish context, the CEFR levels were first mentioned in Andalucía in 2008 in the *Orden* (BOJA 169, 26-08-2008) which determines the curriculum for *Bachillerato*, i.e. the two-year optional secondary education stage before entering University. As stated in the legal document, in those cases in which the student had not taken courses on the second language before, CEFR levels A1 and A2 were determined, whereas CEFR levels A2 and B1 were established when the student had had. In the same line, the guidelines which have been designed and made public by the *Distrito Único Andaluz* since the UEE in 2009-2010 (Distrito Único Andaluz, 2010; 2011) also use the CEFR levels and make explicit that *Bachillerato* (optional secondary education) is supposed to consolidate a competence level CEFR B1 in a second language. To help raters become familiar with the Can-Do statements, the scoring rubrics for the two expected CEFR levels in the students' production in the English exam in the UEE, i.e. A2 and B1, are reproduced in the guidelines.

The level that students need to master before finishing a degree or entering an MA programme has not been fully established yet in the national context. For instance, CEFR level B1 is required as the minimum leaving level to finish the Degree *Maestro en Educación Infantil* and the Degree *Maestro en Educación Primaria* (see Orden ECI/3854/2007 and Orden ECI/3857/2007). The same CEFR level is an entry requirement for the teacher training MA programme for pre-service secondary school teachers, teachers at Official Language Schools or Vocational education. Apart from these two BA degrees and the MA programme, each university decides on the minimum language requirements which students will need to meet to finish a degree, as highlighted in Halbach, Lázaro Lafuente and Pérez Guerra (2010).

In the international context, a very similar picture is found. CEFR level B1 has also been imposed in some countries as a requirement for secondary-school leavers in Chile (Khalifa, Robinson, & Harvey, 2010), and Colombia (Gómez Montes, Mariño, Pike, & Moss, 2010), or is considered a distinguishing characteristic in the CVs of students in China (Xueling, Meizi, & Bateman, 2010). CEFR level B1 is also established as the minimum leaving level (Randall, 2010) for undergraduate students to finish their degree in France and Italy. Finally, other institutions consider that a higher CEFR level, B2 or a C level, is required to enter university, or have a professional career in which English is needed (Green, 2008).

As can be observed, there is still room for debate on the entrance and leaving levels which are to be required at different universities. What seems to be clear is the agreement on the use of the CEFR to promote common ground and standards to establish entrance and leaving requirements, and to design language syllabuses, curricula, examinations, and teaching materials.

Sources of Measurement Errors when Evaluating Writing: the Rating Scale and the Raters

Reliability is a quality of test scores (Bachman, 1990), which refers to the extent to which they result from a test which is free from measurement error, regardless the time when the test is taken, the test form, the raters, etc. (Bachman, 1990; Hamp-Lyons, 1991b; Weigle, 2002; etc.). Due to the importance of reliability in a high-stakes examination such as the UEE, the measurement errors which are caused by the raters and the rating scale when assessing writing will be described.

The factors which trigger the part of the measurement error which stem from the rater have been mentioned in a number of publications (Bachman & Palmer, 1996; Lumley, 2005; McNamara, 1996; Shaw & Weir, 2007; Weigle, 2002; Weir, 2005; etc.), since the reliability of the rating normally refers to the raters' reliability (Hamp-Lyons, 2007). In fact, ensuring that the raters know how to comprehend and apply the scale is considered one out of the two central considerations in scoring, the other being the definition of the rating scale (Weigle, 2002).

Although rating scales or scoring criteria are supposed to help the raters interact with the students' texts and score them in similar ways, the raters' leniency or severity may not change (Kondo-Brown, 2002; McNamara, 1996; Weigle, 1998). In fact, raters are influenced by different factors when rating. For instance, the effect of the use of different scales, their years of rating experience, their academic background, age and gender have been reported (Vaughan, 1991; reviews in Weigle, 2002), with contradictory results on some occasions.Thus, the use of holistic scales by experienced raters seem to result in more lenient scores, while this experience does not have any effect when using analytic scales (Song & Caruso, 1996). However, when experienced raters are compared to less experienced ones, the former are reported to be stricter (Sweedler-Brown, 1985), to use more efficient strategies and a wider knowledge of sources to judge the text (Cumming, 1990), to read the text in one go, to evaluate it by providing comments at the end and by using a limited variety of actions as compared to less expert raters (Huot, 1993; Pula & Huot, 1993; Wolfe & Ranney, 1996). Nevertheless, other studies have reported that inexperienced raters are more severe (Weigle, 1998). Another important aspect is the raters' academic background. For instance, raters who specialise in ESL and other faculty members score essays differently or apply assessment criteria in a different way (Hamp-Lyons, 1991a; Weigle, Boldt, & Valsecchi, 2003). However, it seems that the use of analytic scales diminishes the differences between faculty from different backgrounds (Song & Caruso, 1996). The fact of being a native or non-native speaker of the language evaluated also seems to be a variable to consider, since non-native speakers of the language have been shown to be stricter than native speakers (Bueno González, 1992; Hyland & Anan, 2006). Finally, the rater's age has also been highlighted as a variable which may influence rating, since more leniency is found in older professors on some occasions (Santos, 1988; Vann, Meyer, & Lorenz, 1984), although this may not always be the case (Roberts & Cimasko, 2008).

Apart from the raters' peculiar characteristics, the use of rating scales is a difficult task in itself. As showed by DeRemer (1998) there are different ways how raters proceed when using them. If the rating scale does not provide enough information to describe the students' texts in full or the descriptions in the scale are not clearly specified, the raters are at a loss. Consequently, raters develop strategies to cope with the task, and it may be the case that a similar understanding of the rating category contents may not lead to a similar application of the contents of the scale (Lumley, 2002), and raters may not consider or evaluate the contents of that category in a similar way (Turner & Upshur, 2002).

Other important variables related to the use of a scale are the number of aspects to which the rater has to pay attention (which are included in the scale), or the number of scales a test consists of. The CEFR suggested that four or five categories per level are the limit to avoid that the cognitive load demanded affects raters (Council of Europe, 2001). The number of levels in a scale may also affect the raters' task, since they may not be able to discern between the levels established in the scale (Bachman & Palmer, 1996; Penny, Johnson & Gordon, 2000). For this reason, a reliable and practical number of levels needs to be considered (Bachman & Palmer, 1996). According to Penny, Johnson and Gordon (2000), more than eight levels in a scale may pose problems in inter-rater agreement, and the CEFR advocates for six levels, which are claimed to correspond with the natural levels with which teachers are familiar, namely beginner, elementary, lower intermediate, intermediate, upper intermediate and advanced (Council of Europe, 2001).

Apart from the need for rater training, score resolution methods may be used to increase the raters' reliability when the effects of the previous variables may bias the results. Among the most commonly used ones are those which involve either combining scores from two raters, substituting the score by that provided by an expert, the combination of the scores of the two raters and that of the expert and, finally, the combination of the expert's score with the closest score of one of the original raters (Johnson, Penny, & Gordon, 2000; Weigle, 2002). Another option to improve inter-rater reliability is that raters could augment integer-level scores by adding an additional decimal (Cronbach, Linn, Brennan, & Haertel, 1995, cited in Penny, Johnson and Gordon, 2000). The advantages obtained by doing so include the reduction of disagreement, the increase of inter-rater reliability and the opportunity to allow raters to express the degree of ambiguity found when rating a text. According to a study by Penny, Johnson and Gordon (2000), when using augmented scores, the mean and the standard deviation did not change significantly, but the inter-rater agreement increased (Penny, Johnson, & Gordon, 2000). Finally, scaling, i.e. a statistical analysis which detects raters whose markings are not within the mean and the standard deviation of the raters as a group (Shaw & Weir, 2007), and the development and analysis of automated writing evaluation systems (AWEs) (Burstein & Chodorow, 2002), are also possible.

The effects that the raters' variables may have on their scoring have also been explored in the English exam in the UEE in Spain. For instance, Herrera Soler (2000-2001) addressed the variables of the raters' gender and working place to conclude that males are more lenient than females, and that accuracy was a more important aspect for secondary school teachers than for university female teachers. In another study, Amengual Pizarro (2005) claimed that there were differences between women and men working in high schools, since women were less strict than men. However, the opposite scenario was found when raters work at university, therefore supporting the results by Herrera Soler (2000-2001), but differing from his results that the strictest scoring group is that of women working in high schools (Amengual Pizarro, 2005). The type of evaluation done, i.e. holistic or analytic, also varied if males or females rated the exams. Thus, with holistic ratings and considering the same working place, women awarded higher scores, but the use of analytical evaluation resulted in men scoring higher (Amengual Pizarro, 2005).

The inter-rater agreement found when using holistic or analytic rating for the English exam in the UEE in Spain has also been analysed. Thus, inter-rater agreement with holistic scoring evaluation has been found to be poor (Amengual Pizarro, 2003-2004; Amengual Pizarro & Herrera Soler, 2003), as shown by the intra-class correlation estimate (*k*= .6556) (Amengual Pizarro, 2003). Further studies on the raters' agreement when using analytic and holistic evaluation also showed that the total raters' agreement was not high with holistic evaluation (*k*= .6390), but was slightly higher than that when analytic ratings were used (*k*= .5993) (Amengual Pizarro, 2005). Finally, holistic and focused holistic evaluation have been compared (Watts & García Carbonell, 1998, 2005). The results show that the use of focused holistic criteria (which included six degrees of correctness and a descriptor per degree) provides better results.

Methodology

This section will be divided into two subsections. The first one will describe the English exam in the UEE in the *Distrito Único Andaluz* in 2008, and will explain the selection of the learner corpus in this study. The second subsection will provide brief bios of the two raters who evaluated the students' compositions following the CEFR guidelines, and the way how the rating took place.

The English Exam in the University Entrance Examination: the Selection of the Learner Corpus

The English language exam in 2008 in Andalucía was composed of three main parts, namely *Comprehension, Use of English* and *Production*. For the purposes of this study, only the last part of the exam was considered, i.e. the *Production* section, where students were asked to write 80-100 words on one topic out of the two offered.

2,611 students took the exam for English in the UEE in June 2008 in Jaén. The two topics which were offered to the students on that occasion were «Where, outside Spain, would you like to go on a short pleasure trip?» and «Attracting more tourists is essential for the Spanish economy. Discuss». Since the raters' evaluation of the compositions could be biased by the different genres which these prompts elicit, only the compositions written as an answer to the prompt which was chosen by the highest number of students were selected. In this case, most students (1,406) decided to write on «Where, outside Spain, would you like to go on a short pleasure trip?». Simple random sampling was used (Cochran, 1977) (CI= 95%, p= q = .50) with the program *Stats 1.1*. to find out that 302 compositions were needed to obtain a representative sample of the 1,406 texts written on that topic. Consequently, 302 compositions were randomly selected and composed the learner corpus analysed in this paper. The students' pieces of writing were then converted into electronic format, paying special attention to keep a faithful transcription of the students' compositions, and the total amount of words of the learner corpus used, i.e. 34,403, was calculated.

The Raters and the Rating Process

Two female raters, who are native speakers of English, were selected to evaluate each of the students' compositions in the learner corpus. Since the raters' backgrounds prove decisive when undertaking the rating process, brief bios will be provided here.

Rater 1 speaks French and German as FLS, and is a bilingual speaker of Spanish, since she has been living in Spain for more than 32 years now. She teaches in a secondary school, she also teaches the course «That's English» at the Official Language School in Jaén, and offers private tuition at various proficiency levels. Rater 2 speaks French as an FL and is also a bilingual speaker of Spanish. She has taught English oral classes in courses whose level range from A2 to C1 for three academic years at various departments at the *Universidad de Murcia*, where she has also been involved in various research projects related to the use of English as an FL.

Although both raters use the CEFR due to their jobs, following Salamoura's (2008) recommendation, they were instructed to refer to the materials which are available, namely the CEFR scales for written production (Council of Europe, 2001), and Table 5.8 entitled «Written Assessment Criteria Grid» in the manual for *Relating Language Examinations to the Common European Framework of Reference* (2003),[2] for them to double check that they were applying the criteria correctly.

In the grading process, rating augmentation (Cronbach, Linn, Brennan, & Haertel, 1995) was allowed at all levels, even though the CEFR only allows rater augmentation at levels A2, B1 and B2. Furthermore, the use of the minus bands was also permitted, since the raters felt it was necessary for a better grading. Therefore, when a rater felt that a paper was below the standards of a level (e.g. B1), but was not to be considered within the lower augmentated rater level (A2+), she was allowed to use a B1- level (see Table I).

The flexibility of use of the CEFR levels made it possible to reduce the number of levels from 18 (when using rating augmentation and the minus bands) to the basic 3 levels, as needed for the analyses. Thus, if a reduction from 18 to 12 levels was needed, it could be done by including those compositions which were awarded an A1- and an A1 into a broader A1 level (and the same with A2- and A2, B1- and B1, B2- and B2, C1- and C1, or C2- and C2). Similarly, the 12 levels could be also included within 6 levels, so that A1 and A1+ levels may be considered an A1 level. Finally, the 6 levels could also be reduced to 3 broad levels, if A1 and A2 are included within the general A level.

TABLE I. Flexibility of the CEFR levels

CEFR: three levels	CEFR: six levels	CEFR: twelve levels	CEFR: eighteen levels
A Basic User	A1 Breakthrough	A1	A1-
			A1
		A1+	A1+
	A2 Waystage	A2	A2-
			A2
		A2+	A2+

[2] A more recent version of the *Written Assessment Criteria Grid* is found in table C-4 (Council of Europe, 2009).

B Independent User	B1 Threshold	B1	B1-
			B1
		B1+	B1+
	B2 Vantage	B2	B2-
			B2
		B2+	B2+
C Proficient User	C1 Effective Operational Proficiency	C1	C1-
			C1
		C1+	C1+
	C2 Mastery	C2	C2-
			C2
		C2+	C2+

To consider the CEFR levels for statistic analyses each CEFR (when considering 3, 6, 12 or 18 levels) was given a number, as seen in Table II. To do so, four scales were designed. The first scale ranges from 1 to 3 (for 3 CEFR levels), the second one from 1 to 6 (for the 6 CEFR levels), the third one from 1 to 12 (when considering augmentation, i.e. 12 CEFR levels), and the last scale from 1 to 18 (when augmentation and the minus bands were considered, i.e. 18 CEFR levels).

TABLE II. Conversion of CEFR levels into numerical data

Three levels		Six levels		Twelve levels		Eighteen levels	
Level	Number	Level	Number	Level	Number	Level	Number
A	1	A1	1	A1	1	A1-	1
						A1	2
				A1+	2	A1+	3
		A2	2	A2	3	A2-	4
						A2	5
				A2+	4	A2+	6
B	2	B1	3	B1	5	B1-	7
						B1	8
				B1+	6	B1+	9
		B2	4	B2	7	B2-	10
						B2	11
				B2+	8	B2+	12

C	3	C1	5	C1	9	C1-	13
						C1	14
				C1+	10	C1+	15
		C2	6	C2	11	C2-	16
						C2	17
				C2+	12	C2+	18

Results and Discussion

The Use of the CEFR: Inter-realiability

After each CEFR level was related to a number on a scale, as indicated in Table II, the distance between the grade awarded by one rater and the other could be calculated. This was made as follows. If the 3 broad levels were taken into account (A, B and C), the distance between one rater's A1 and the other rater's B1, awarded to the same composition, would be 1. If 6 levels were taken into account, the distance between A1 and B1 would be 2. However, when 12 CEFR levels were employed, the distance would be 4, and with the use of 18 CEFR levels the distance would be 6.

Once the distances were found out, the analysis of such data revealed that, when using 18 CEFR levels both raters awarded the same CEFR level to the same composition in 47 cases, and the mode in the breach of agreement was 105 cases (i.e. 34.8% of the total number of cases), involving a difference of only one level in the raters' opinion. In other words, a rater awarded B1 to a composition, whereas the other one awarded B1+ or B1-, or a rater awarded C1- and the other one B2+ or C1. The data in Table III below also shows that the major difference in the raters' levels awarded to the same composition was seven, which occurred on two occasions only. More specifically, in one composition a rater awarded A1- to a composition, whereas the other opted for B1 level, and in the second composition, a rater provided A1, whereas the other one awarded B1+.

TABLE III. Distance when using 18 CEFR levels

Valid	Frequence	Percentage	Valid Percentage	Accumulated percentage
0	47	15.6	15.6	15.6
1	105	34.8	34.8	50.3
2	55	18.2	18.2	68.5
3	35	11.6	11.6	80.1
4	39	12.9	12.9	93.0
5	7	2.3	2.3	95.4
6	12	4.0	4.0	99.3
7	2	.7	.7	100.0
Total	302	100.0	100.0	

If 12 levels were considered, Table IV shows that there were 100 cases in which raters agreed, and 107 compositions in which the two raters differed in one level when awarding a CEFR level to each composition (i.e. 35.4% of the total number of cases). For instance, if a rater awarded A2 level, the other rater awarded either A2+ or A1+ level. When using 12 levels, the major difference was found in one case in which raters differed in five levels, i.e. one rater awarded A1 and the other assigned B1+, but it is also worth mentioning that in thirteen compositions (4.3% of the cases), raters' differed in 4 levels.

TABLE IV. Distance when using 12 CEFR levels

Valid	Frequence	Percentage	Valid percentage	Accumulated percentage
0	100	33.1	33.1	33.1
1	107	35.4	35.4	68.5
2	59	19.5	19.5	88.1
3	22	7.3	7.3	95.4
4	13	4.3	4.3	99.7
5	1	.3	.3	100.0
Total	302	100.0	100.0	

The use of 6 CEFR levels (Table V) reveals that the mode is the agreement of the two raters (in 196 cases, i.e. 64.9% of the total sample), and the major distance between two CEFR levels awarded to the same composition is 2, which happens in 15 cases. In fact, in 14 cases, the distance was that between A1 and B1, whereas in another case the distance was the one between A2 and B2.

TABLE V. Distance when using 6 CEFR levels

Valid	Frequence	Percentage	Valid Percentage	Accumulated Percentage
0	196	64.9	64.9	64.9
1	91	30.1	30.1	95.0
2	15	5.0	5.0	100.0
Total	302	100.0	100.0	

Finally, when considering 3 CEFR levels (Table VI), raters agreed on 226 cases. Even when using the three basic levels, there were 76 occasions (25.2% of the total amount of compositions) on which they differed when awarding A or B to the same composition.

TABLE VI. Distance when using 3 CEFR levels

Valid	Frequence	Percentage	Valid Percentage	Accumulated Percentage
0	226	74.8	74.8	74.8
1	76	25.2	25.2	100.0
Total	302	100.0	100.0	

As a summary, the information obtained regarding the raters' agreement when awarding a CEFR level to the 302 compositions can be summarized as follows:

TABLE VII. Inter-rater agreement when using 3, 6, 12 and 18 CEFR levels

	18 CEFR levels	12 CEFR levels	6 CEFR levels	3 CEFR levels
Inter-rater Agreement	47 cases (15.56%)	100 cases (33.11%)	196 cases (64.9%)	226 cases (74.83%)

As can be seen, the number of occasions on which the raters did not coincide when awarding a level to a composition was high, even when considering 3 levels (25.17%). When using 6 CEFR levels, there are 15 cases in which the raters differ in two levels when awarding a grade, and there are 91 cases in which they award a grade which differs in one level (see Table 5). In other words, there are 30.1% of the cases in which there are problems to distinguish between CEFR level A1 and A2, or A2 and

B1, or B1 and B2. Therefore, the functional rubrics provided for those levels seem not to be clear enough for raters to grade 25.17% of the compositions in the case of the use of 3 CEFR levels, and 35.1% when using 6 CEFR levels. As claimed by Alderson et al (2004), Huhta et al. (2002) or Kaftandjieva and Takala (2002), the scales may need further description so that they may be easily understood and applied.

The results obtained when using 12 or 18 CEFR levels, i.e. rating augmentation and the use of the minus bands, highlight the decrease of the raters' total agreement, contrary to the results in Penny, Johnson and Gordon (2000) when using augmentation in holistic scoring. In fact, the use of more CEFR levels reduces the percentage of total raters' agreement in half of the percentage, a finding in line with Penny, Johnson and Gordon's (2000) claim that using more than 8 levels on a scale poses limitations to the inter-rater agreement.

The use of the minus bands, as suggested by the raters' claim that the addition of that band would improve their grading, entailed even lower inter-rater agreement. As seen in the difference found when using 12 and 18 CEFR levels, the percentages of compositions which were graded with the same CEFR level were 33.11% and 15.56%, respectively. The use of the minus bands was then found not to improve the raters' total agreement, but to reduce it, since less than half the percentage of compositions received the same CEFR level when using it. Apart from this result, no compositions were placed by both raters at any of the minus bands, which indicates that its use may not be recommended.

The inter-raters' reliability was then calculated by using Cohen's kappa coefficient. The low inter-rater agreement obtained (*k*= .245),[3] even lower than those reported by Amengual Pizarro (2003, 2005), indicates that it is difficult for raters to discern which level a composition is to be awarded, a caution which has been previously raised by Bachman and Palmer (1996) and Penny Johnson and Gordon (2000). This is specially so when raters do not know how to use the scale (Weigle, 2002), when there are no benchmark papers or when the information provided in the rubric is not clear or enough for raters to decide. The use of the 6 CEFR levels advocated for the CEFR which, besides, correspond with the natural levels with which teachers (Council of Europe, 2001) and experts are familiar, seems to be the best option when trying to find a balance between the number of levels used and the inter-raters' agreement, thus obtaining a more reliable grading in a high-stakes examination such as the English exam in the UEE.

[3] According to Altman (1991), values above k= .8 are indicators of a very good agreement.

The effects exerted by the raters' personal characteristics and the grades that they awarded also revealed interesting findings. Although the gender variable was considered when choosing the raters who graded the students' compositions (both of them are females), their age, teaching background and experience was different. Whereas rater 1 has a wide teaching experience at secondary school level in Spain, rater 2 has less teaching experience in general and, when teaching in institutions, i.e. not undertaking private tuition, she has mostly worked at university level. As indicated in the literature review, the raters' teaching experience and the institution where they work play a crucial role in the inter-rater reliability found in the grades awarded. In fact, rater 2 provided better grades than rater 1, all in all, so the use of the same scale did not homogenize their grades, as reported by McNamara (1996), Weigle (1998) or Kondo-Brown (2002). In fact, the results obtained with the grades provided by the raters are in line with Sweedler-Brown's (1985) claim that experienced raters are stricter, and with Roberts and Cimasko's (2008) claim that older professors may not be the most lenient ones. Similarly, the working place also seems to play an important role, since the rater who works at secondary education provided students with stricter CEFR grades as compared to rater 2, a finding which is in line with Herrera Soler (2000-2001), but not with Amengual Pizarro (2005).

An Overview of the CEFR Levels Awarded to the Students' Compositions

Due to the attested low inter-rater reliability, and in order to use the most reliable data when analysing the number of compositions at each CEFR level, only those compositions in which there was 100% of inter-rater agreement were considered to analyse the levels which were awarded to the students' compositions in the English exam in the UEE.

As seen in Table VIII below, when considering 18 CEFR levels (i.e. 47 compositions), 1 composition was awarded A2, 23 compositions received B1, 19 compositions were awarded B1+, and 4 compositions obtained B2, while the use of 12 levels classified the 100 compositions into 1 at A2 level, 75 at B1 level, 19 at B1+ level, 4 at B2 level and 1 at C1 level. When using 6 levels (196 compositions), 10 compositions were classified at A2 level, 179 at B1 level, 6 at B2 level and, finally, one composition was included in the C1 level. Finally, the use of 3 levels divided the 226 compositions into 17 at A level, 208 at B level and only 1 at C level.

Therefore, the written production by the students in the learner corpus is characterized by a large number of compositions at B level (208), B1 level (179), B1 level (75), or B1 level (23), considering 3, 6, 12 and 18 levels, respectively. In other words, if 3 levels are considered, 92.03% of the compositions are placed at B level; when considering 6 levels, 94.43% are at B level (either B1 or B2); when using 12 levels, 94% are at B level (either B1, B1+ or B2); and, finally, when using 18 levels, 97.87% are at B level (either B1-, B1, B1+, B2-, B2 or B2+).

TABLE VIII. Inter-rater agreement on the students' compositions, depending on the CEFR levels considered

CEFR: Three Levels	Inter-rater Agreement (no. of cases)	CEFR: Six Levels	Inter-rater Agreement (no. of cases)	CEFR Twelve Levels	Inter-rater Agreement (no. of cases)	CEFR: Eighteen Levels	Inter-rater Agreement (no. of cases)
A Basic User	17	A1 Breakthrough		A1		A1-	
						A1	
				A1+		A1+	
		A2 Waystage	10	A2	1	A2-	
						A2	1
				A2+		A2+	
B Independent User	208	B1 Threshold	179	B1	75	B1-	
						B1	23
				B1+	19	B1+	19
		B2 Vantage	6	B2	4	B2-	
						B2	4
				B2+		B2+	

C **Proficient User**	1	C1 Effective Operational Proficiency	1	C1	1	C1-	
						C1	
				C1+		C1+	
		C2 Mastery		C2		C2-	
						C2	
				C2+		C2+	

If the differences between B, B1 and B1+ levels are considered, the data obtained show that, when using 3 cefr levels, compositions at B level amounted to 92.03% of the total amount of the compositions, and when considering 6 cefr levels, 91.33% were awarded B1 level. In the case of 12 cefr levels, B1 compositions, i.e. B1 and B1+ compositions, amounted to 94% (B1 compositions being 75% of the total amount of compositions, and B1+ compositions, 19%). Finally, the use of 18 cefr levels revealed that 89.36% of the compositions were at B1 level (B1 compositions being 48.94% of the compositions, and B1+ compositions amounting for 40.42% of the total number of compositions).

In the case of those compositions which were not awarded a cefr level B1, it was at levels A (specially A2) and B2 that the compositions were located on the scale. If 3 cefr levels were employed, 7.52% of the compositions are at A level, whereas 0.44% are at C level. In the case of 6 cefr levels, 5.10% of the compositions are awarded A2 level, 3.06% B2 level and 0.51% C1 level. Finally, the use of 12 or 18 cefr levels reveals that the second most frequent cefr levels awarded to compositions are B2 levels, with 4% and 8.51% of the compositions, respectively, and that only one composition was awarded A2 level. As a result, two main observations can be made: first, a very low percentage of compositions is placed at C level (in fact only one composition); and second, the more cefr levels used, the more the raters coincide in the grades awarded to compositions at B levels, if compared to the cases of total agreement in A level compositions (see Table viii).

Conclusions

The main objective of this paper was to use the CEFR for a twofold purpose. First, to explore the CEFR levels which would be awarded to a representative sample of the compositions written for the English exam in the UEE in Jaén in 2008, and see if they met the established levels in the legal documents in Andalucía (BOJA 169, 26-08-2008). Second, to become aware of the main problems encountered when raters use the CEFR and the Can-Do statements to assess the students' compositions in the English exam.

As seen in the results obtained in this paper, the use of the CEFR levels and the Can-Do statements pose some problems to raters who work with the CEFR guidelines. In fact, the low inter-rater agreement obtained even when considering the three basic levels A, B, and C, may indicate that the rubrics are not enough for raters to discern the level to be awarded to compositions. Therefore, a previous step is to be taken if the CEFR levels and the Can-Do statements are to be applied in the assessment of the students' compositions in the English exam. Although rater training may improve the inter-rater reliability, the complementation of the existing rubrics with further descriptions on the students' use of the language expected at each level would be crucial for the raters' better understanding and use of the CEFR.

Apart from the need to complement the rubrics, the results in this paper point to some recommendations regarding the use of rating augmentation and the minus bands, i.e. 12 and 18 levels, respectively. As discussed above, their use is not useful, since the raters' total agreement decreases significantly when they use them, and raters do not agree on the use of the minus band in any case. For this reason, the use of 6 CEFR levels is advocated for if a balance between the number of levels used and the total inter-rater agreement is sought.

The raters' age, teaching background and experience have been found to play a very important role in their rating task, in line with previous publications. In fact, the experienced (and older) rater who works at a secondary school proved to be stricter than the less experienced one, whose teaching experience is at university level. Therefore, these variables need to be considered if biases are to be avoided in the rating process.

Once the compositions to which both raters awarded the same CEFR level were selected (when using 3, 6, 12 and 18 CEFR levels), the findings revealed that most compositions were placed at CEFR B1 level. For instance, the consideration of 6 CEFR levels indicates that 91.33% of the compositions on which there was total inter-rater agreement were placed at B1 level. Therefore, the students show the required competence level at the end of their optional secondary education and in the English exam in the UEE in Andalucía. As expected in the legal documents, the levels which are mainly represented in the exam are A2 and B1 levels, although compositions showing a higher command of the language are also found in the learner corpus.

The main limitation of this study is that the results obtained can only be generalized to the students who took the English exam in the UEE in June 2008 in Jaén, and wrote on a specific topic. Therefore, further research would be needed to replicate this study with the compositions written by other students in Andalucía and other universities in Spain, writing on other topics, and being evaluated by other raters.

Bibliographical References

Altman, D. G. (1991). *Practical Statistics for Medical Research*. London: Chapman and Hall.

Alderson, J. C. (1991). Bands and Scores. In J.C. Alderson & B. North (Eds.), *Language testing in the 1990s,* 71-86). London: Macmillan.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2004). *The Development of Specifications for Item Development and Classification within the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Reading and Listening. Final report of the Dutch construct project*. Unpublished document.

Amengual Pizarro, M. (2003). A Study of Different Composition Elements that Raters Respond to. *Estudios Ingleses de la Universidad Complutense,* 11, 53-72.

— (2003-2004). Rater Discrepancy in the Spanish University Entrance Examination. *Journal of English Studies,* 4, 23-36.

— (2005). Posibles sesgos en los resultados del examen de selectividad. In H. Herrera Soler & J. García Laborda (Ed.), *Estudios y criterios para una selectividad de calidad en el examen de Inglés* (121-148). Valencia: Universidad Politécnica de Valencia.

Amengual Pizarro, M., & Herrera Soler, H. (2003). What is that Raters are Judging? In G. Luque Agulló, A. Bueno González & G. Tejada Molina (Eds.), *Las lenguas en un mundo global* (11-18). Jaén: Servicio de Publicaciones de la Universidad de Jaén.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bueno González, A. (1992). Errores en la elección de palabras en inglés por alumnos de Bachillerato y cou. In A. Bueno González, J. A. Carini Martínez & A. Linde López (Eds.), *Análisis de errores en inglés: tres casos prácticos* (39-105). Granada: Servicio de Publicaciones de la Universidad de Granada.

Burstein, J. & Chodorow, M. (2002). Directions in Automated Essay Analysis. In R. B. Kaplan (Ed.), *The Oxford Handbook of Applied Linguistics* (487-497). New York: Oxford University Press.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

— (2003). *Relating Language Examinations to the Common European Framework of Reference for Languages:Learning, Teaching, Assessment, Manual:Preliminary Pilot Version*. Strasbourg: Council of Europe, Language Policy Division.

— (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (cefr). A Manual*. Strasbourg: Council of Europe, Language Policy Division.

Cochran, W. G. (1977). *Técnicas de muestreo*. México. Trillas

Cronbach, L. L., Linn, R., Brennan, R. & Haertel, E. (1995). *Generalizability Analysis for Educational Assessments*. Los Angeles: Center for the Study of Evaluation, Standards, and Student Testing, University of California at Los Angeles.

Cumming, A. (1990). Expertise in Evaluating Second Language Compositions. *Language Testing,* 7, 31-51.

DeRemer, M. L. (1998). Writing Assessment: Raters' Elaboration of the Rating Task. *Assessing Writing,* 5, 7-29.

Díez-Bedmar, M. B. (in press). Spanish Pre-university Students' use of English: cea Results from the University Entrance Examination. *International Journal of English Studies,* 11.

Díez-Bedmar, M. B. (in press). The English Exam in the University Entrance Examination: an Overview of Studies. *Revista Canaria de Estudios Ingleses,* 63.

García Laborda, J. (2006). Analizando críticamente la selectividad de Inglés ¿Todos los estudiantes españoles tienen las mismas posibilidades? *Tesol Spain,* 30, 9-12.

Gómez Montes, I., Mariño, J., Pike, N. & Moss, H. (2010). Colombia National Bilingual Project. *Research Notes,* 40, 17-22.

Green, A. (2008). English Profile: Functional Progression in Materials for elt. *Research Notes,* 33, 19-25.

Hamp-Lyons, L. (1991a). Scoring Procedures for esl Contexts. In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (241-276). Norwood, NJ: Ablex Publishing Corporation.

— (1991b). Basic Concepts. In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (5-15). Norwood NJ: Ablex.

— (2007). Editorial: Worrying about Rating. *Assessing Writing,* 12, 1-9.

Herrera Soler, E. (2000-2001). The Effect of Gender and Working Place of Raters on University Entrance Examination Scores. *Revista Española de Lingüística Aplicada,* 14, 161-180.

Huhta, A., Luoma, S., Oscarson, M., Sajavaara, K., Takala, S. & Teasdale, A. (2002). A Diagnostic Language Assessment System for Adult Learners. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Case Studies* (130-146). Strasbourg: Council of Europe.

Huot, B. A. (1993). The Influence of Holistic Scoring Procedures on Reading and Rating Student Essays. In M. M. Williamson, & B.A. Huot (Eds.), *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations* (206-236). Cresskill, NJ: Hampton Press.

Hyland, K. & Anan, E. (2006). Teachers' Perceptions of Errors: the Effects of First Language and Experience. *System,* 34, 509-519.

Johnson, R. L., Penny, J. & Gordon, B. (2000). The Relation between Score Resolution Methods and Interrater Reliability: an Empirical Study of an Analytic Scoring Rubric. *Applied Measurement in Education,* 13, 121-138.

Kaftandjieva, F. & Takala, S. (2002). Council of Europe Scales of Language Proficiency: a Validation Study. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Case Studies* (106-129). Strasbourg: Council of Europe Publishing.

Khalifa, H., Robinson, M. & Harvey, S. (2010). Working Together: the Case of the English Diagnostic Test and the Chilean Ministry of Education. *Research Notes,* 40, 22-26.

Kondo-Brown, K. (2002). A facets Analysis of Rater Bias in Measuring Japanese L2 Writing Performance. *Language Testing,* 19, 3-31.

Lumley, T. (2002). Assessment Criteria in a Large-scale Writing Test: What Do they Really Mean to the Raters? *Language Testing,* 19, 246-276.

— (2005). *Assessing Second Language Writing: the Rater's Perspective*. Frankfurt: Peter Lang.

McNamara, T. (1996). *Measuring Second Language Performance*. London: Longman.

Fecha de entrada: 10-07-2011 Fecha de aceptación: 25-10-2011

Penny, J., Johnson, R. L. & Gordon, B. (2000). The Effect of Rating Augmentation on Inter-rater Reliability. An Empirical Study of a Holistic Rubric. *Assessing Writing,* 7, 143-164.

Pula, J. J. & Huot, B. A. (1993). A Model of Background Influences on Holistic Raters. In M. Williamson & B. Huot (Eds.), *Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations* (237-265). Cresskill, NJ: Hampton Press.

Randall, S. (2010). Cambridge esol's Growing Impact on English Language Teaching and Learning in National Education Projects. *Research Notes,* 40, 2-3.

Roberts, F. & Cimasko, T. (2008). Evaluating esl: Making Sense of University Professors' Responses to Second Language Writing. *Journal of Second Language Writing,* 17, 125-143.

Salamoura, A. (2008). Aligning English Profile Research Data to the cefr. *Research Notes,* 33, 5-7.

Santos, T. (1988). Professors' Reactions to the Academic Writing of Nonnative Speaking Students. *tesol Quarterly,* 22, 69-90.

Shaw, S. D. & Weir, C. J. (2007). *Examining Writing. Research and Practice in Assessing Second Language Writing*. Cambridge: Cambridge University Press.

Song, B. & Caruso, I. (1996). Do English and esl Faculty Differ in Evaluating the Essays of Native English-Speaking, and esl Students? *Journal of Second Language Writing,* 5, 163-182.

Sweedler-Brown, C. O. (1985). The Influence of Training and Experience on Holistic Essay Evaluation. *English Journal,* 74, 49-55

Turner, C. E. & Upshur, J. A. (2002). Rating Scales Derived from Student Samples: Effects of the Scale Maker and the Student Sample on Scale Content and Student Scores. *tesol Quarterly,* 36, 49-70.

Vann, R. J., Meyer, D. E. & Lorenz, F. O. (1984). Error Gravity: a Study of Faculty Opinion of esl Errors. *tesol Quarterly,* 18, 427-440.

Vaughan, C. (1991). Holistic Assessment: What Goes on in the Rater's Mind? In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (11-125). Norwood, NJ: Ablex.

Watts, F. & García Carbonell, A. (2005). Control de calidad en la calificación de la prueba de lengua inglesa de Selectividad. In H. Herrera Soler & J. García Laborda (Eds.), *Estudios y criterios para una selectividad de calidad en el examen de inglés* (99-115). Valencia: Universidad Politécnica de Valencia.

Weigle, S. C., Boldt, H. & Valsecchi, M. I. (2003). Effects of Task and Rater Background in the Evaluation of esl Student Writing: a Pilot Study. *tesol Quarterly,* 37, 345-354.

Weigle, S. C. (1998). Using facets to Model Rater Training Effects. *Language Testing,* 15, 263-287.

— (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (2005). Limitations of the Common European Framework for Developing Comparable Examinations and Tests. *Language Testing,* 22, 281-300.

Wolfe, E. W. & Ranney, M. (1996). Expertise in Essay Scoring. In D. C. Adelson & E. A. Domeshek (Eds.), *Proceedings of icls 96* (545-550). Charlottesville, VA: Association for the Advancement of Computing in Education.

Xueling, C., Meizi, H. & Bateman, H. (2010). The Use of bec as a Measurement Instrument in Higher Education in China. *Research Notes,* 40, 13-15.

Electronic Resources

Distrito Único Andaluz (2010). *Directrices y orientaciones generales para las pruebas de acceso a la universidad. Curso 2009-10. Lengua extranjera inglés*. Retrieved on June, 16th 2010, from: http://www.ujaen.es/serv/acceso/documentos/orient_selectiv_2009_2010/ingles.pdf.

— (2011). *Directrices y orientaciones generales para las pruebas de acceso a la universidad. Curso 2010-11. Lengua extranjera inglés*. Retrieved on June, 15th, 2011, from: http://www.ujaen.es/serv/acceso/documentos/orient_selectiv_2010_2011/ingles.pdf.

Halbach, A., Lázaro Lafuente, A. & Pérez Guerra, J. (2010). *La acreditación del nivel de lengua inglesa en las universidades españolas*. British Council. Retrieved on July, 14th 2011, from: http://www.britishcouncil.org/spain_informe_acreditacion_ingles_universidades_espanolas_.pdf.

Orden eci/3854/2007, de 27 de diciembre, por la que se establecen los requisitos para la verificación de los títulos universitarios oficiales que habiliten para el ejercicio de la profesión de Maestro en Educación Infantil. *Boletín Oficial del Estado,* 312. Retrieved on May, 20th, 2008, from: http://www.boe.es/boe/dias/2007/12/29/pdfs/A53735-53738.pdf.

Orden eci/3857/2007, de 27 de diciembre, por la que se establecen los requisitos para la verificación de los títulos universitarios oficiales que habiliten para el ejercicio de la profesión de Maestro en Educación Primaria. Boletín Oficial del Estado, 312.

Retrieved on May, 20th, 2008, from: http://www.boe.es/boe/dias/2007/12/29/pdfs/A53747-53750.pdf

Watts, F. & García Carbonell, A. (1998). Rater Agreement in English Language Assessment in the Spanish University Aexamination Battery. *Language Testing Update* 23, Retrieved on November, 10th, 2008, from: http://www.upv.es/diaal/publicaciones/watts3.pdf.

Dirección de contacto: María Belén Díez-Bedmar. Edificio D-2. Departamento de Filología Inglesa. Facultad de Humanidades y Ciencias de la Educación. Universidad de Jaén. Paraje las Lagunillas. 23071, Jaén. E-mail: belendb@ujaen.es

Fecha de entrada: 10-07-2011 Fecha de aceptación: 25-10-2011