A Study of Different Composition Elements that Raters Respond To

Estudio de las respuestas de los evaluadores a diferentes elementos de las redacciones

Marian Amengual Pizarro

Universitat de les Illes Balears (UIB)
dfemap0@ps.uib.es

Recibido: 24 de febrero de 2003
Aceptado: 8 de abril de 2003

ABSTRACT

This study investigated the reactions of thirty-two raters, not trained in ESL evaluation techniques, to three University Entrance Examination compositions representative of three different score levels of ESL proficiency (i.e. low, middle and high) . Raters were asked to evaluate compositions holistically. They were also asked to indicate the best and worst features of each composition and to relate them to the following categories: *content, organization, grammar, vocabulary, register, mechanics* and *presentation*. Finally, raters were instructed to judge a list of fourteen sentences, each containing one of seven error types associated with the previously categorised analytic features. The results were as follows: a) holistic scores showed a substantial discrepancy across raters; b) raters were influenced by salient features of the compositions; c) grammar was identified as a primary positive and negative feature in the final judgement of the compositions; d) raters adjusted their marking behaviour to the proficiency level of the compositions and e) raters showed a great variability in attention and importance attached to different criteria.

KEY WORDS

Evaluation of compositions. University Entrance Examination. Evaluation features. ESL.

RESUMEN

Este trabajo investigó las reacciones de treinta y dos calificadores, sin formación específica en técnicas de evaluación de inglés como segunda lengua, ante tres redacciones de las Pruebas de Acceso a la Universidad, correspondientes por la nota a tres niveles de destreza del inglés como segunda lengua (bajo, medio y alto). Se pidió a los calificadores que evaluaran las composiciones globalmente. También se les pidió que indicasen los rasgos mejores y peores de cada redacción, y que los asignaran a las categorías siguientes: *contenido, organización, gramática, vocabulario, registro, mecánica* y *presentación*. Finalmente, los calificadores recibieron la instrucción de juzgar una lista de catorce oraciones, todas las cuales contenían uno de los siete tipos de errores asociados con los rasgos analíticos que antes habían asignado a las categorías mencionadas arriba. Los resultados fueron los siguientes: a) las notas globales mostraron discrepancias sustanciales según los calificadores; b) los calificadores se vieron influidos por los rasgos prominentes de las composiciones; c) la gramática se identificó como un rasgo primordial positivo o negativo en el juicio final sobre las composiciones; d) los calificadores ajustaron su puntuación al nivel de destreza de las composiciones y e) los calificadores mostraron gran variabilidad en la atención y la importancia concedida a los diferentes criterios.

PALABRAS CLAVE

Evaluación de redacciones. Prueba de Acceso a la Universidad. Rasgos evaluables. Inglés como segunda lengua.

1. Introduction

One of the key issues when researching ESL evaluation is to identify those aspects of student performance that lead to academic success in writing. More specifically, researchers are interested in the features that raters focus on when evaluating ESL students' compositions.

Holistic scoring is based on *subjective* assessments and this is used to make important decisions in students' academic lives. Its validity as a rating method has been questioned by some researchers who aim to strike a balance between pure subjectivity and objective precision (Huot 1990; Hamp-Lyons 1991; Vaughan 1991).

Abundant literature exists addressing the analysis of the dominant aspects of student performance that influence raters' judgements but the results obtained have been contradictory. Freedman (1979), for example, identifies content as the major factor that influences raters' scores, while Santos (1988) concludes that raters are primarily affected by lexical errors. Other studies have found that English language proficiency, especially the absence of error. is foremost in the evaluation of ESL writing (Mullen 1980; Homburg 1984; McDaniel, 1985; Sweedler-Brown 1993). Researchers such as Oller (1979), however, maintain that the overriding criterion in writing should be to judge communicative effectiveness. More recently, Dordick (1996) calls for raters to look at errors that interfere most seriously with comprehension. This position is opposed to other studies such as Vann, Meyer, and Lorenz (1991) which seem to be more interested in measuring raters' personal reaction to student errors.

It may be true that different factors affect raters when they assign global scores to students' compositions. Since a comparison of holistic scores would not provide information on the relative weight that raters attach to different criteria, we believe that an analysis of the criteria raters use when looking at the same student compositions would be more sensitive to such differences.

Taking a closer look at the raters' criteria, this study aims to investigate the components in students' compositions that influence raters' scores as well as the subjective reaction to errors in ESL students' writing. The use of methodologies based on the analysis of isolated sentences presenting error samples has been criticised by some researchers who question the validity of such procedures (Davies 1983). Thus, raters in this study were asked to respond to both error samples in individual sentences as well as errors presented in the text.

2. Purpose

The central purpose of this study aims to investigate the extent of agreement of ESL raters in assessing the same set of compositions. It also attempts to investigate which features of the text most strongly influence the global assessment of these compositions and the raters' judgements of, and personal reactions to errors and the language produced in students' writing. To that end, the following research questions were posed:

1. Are raters significantly different in their holistic evaluations of ESL composition samples?
2. Which composition elements influenced raters the most? Which errors appeared most salient to the raters who read and corrected the compositions?
3. What is the relationship between the raters' judgements of errors and their final scores?

3. Method

For the purpose of this study three English compositions from the University of the Balearic Islands Entrance Examination were set aside and independently assessed by thirty-two raters working in University and Secondary Education. The raters were qualified English Language teachers and they had all participated in the assessment of the English Test in the University Entrance Examination, which took place in Madrid and Palma de Mallorca in June 2000.

The compositions were chosen from a pool of fifty tests, on the basis of their overall means, as representative of three different score levels of ESL proficiency: low ($\bar{X}$ = 2.47), middle ($\bar{X}$ = 4.56) and high ($\bar{X}$ = 7.78). The actual composition topic was "A holiday in London mentioning the places you would visit and why."

Each rater was asked to read and rate the three compositions holistically on a scale of 1 to 10 as they would ordinarily rate compositions in the English Test. Since in the University Entrance Examination subtests there are no clearly defined guidelines for answers (Herrera 1999), no marking scheme was deliberately specified to ensure normal subjective impressionistic marking. In doing so, our purpose was to discover raters' perspectives about what constitutes good performances of L2 writing. The raters were asked to read the compositions in the same order (i.e. first low, second middle and third high) so as to eliminate the possible effects due to order of presentation.

After scoring each composition, all raters were asked to note down the best and worst features of each composition and place them in the following categories: *content*, *organization*, *grammar*, *vocabulary*, *register*, *mechanics* and *presentation*. The purpose of this phase was to identify the influence of different composition elements on ratings as well as to consider the relationship between individual scores and judgements.

In addition to the above rating assignment, all raters were asked to judge a list of fourteen original sentences, each containing one of seven error types commonly committed by ESL students, and which had been associated with the previous categorised analytic features (i.e. *content*, *organization*, *grammar*, *vocabulary*, *register*, *mechanics* and *presentation*). Raters were instructed to judge the severity of the errors on a 5-point scale of acceptability or tolerance, from *not at all important* (1) to *very important* (5). To aid raters in their task, errors were highlighted.

Therefore, the three-part data collection session was as follows:

1) A comparison between individual rater's holistic scores;
2) An examination of the influence of different composition elements on ratings and an analysis of the seriousness of particular errors presented in the original students' compositions as well as in isolated sentences;
3) An examination of the relationship between judgements and scores of individual raters.

4. Results and discussion

4.1. *Holistic scores*

Table 1 shows the frequency distribution of the individual scores assigned by the thirty-two raters to the lowest, middle and highest ranked compositions. A simple comparison was also made of the holistic scores given to the same compositions by the same raters. Our first aim here was to examine inconsistencies in holistic scores. The range of composition quality, the means and standard deviations of compositions as scored by raters are presented in Table 2.

Table1. Frequency distribution of the scores awarded by the 32 raters on the lowest, middle and highest scored compositions.

Scores	Lowest ranked composition (frequency)	Middle ranked composition (frequency)	Highest ranked composition (frequency)
1	6	0	0
2	12	3	0
3	9	5	0
4	4	7	0
5	0	9	5
6	1	5	3
7	0	2	2
8	0	1	9
9	0	0	10
10	0	0	3
TOTAL	32	32	32

Table 2. Descriptive Statistics for Raters and Compositions

	N	Range	Minimum	Maximum	Mean	Std. Deviation
Lowest ranked composition	32	5	1	6	2.47	1.14
Middle ranked composition	32	6	2	8	4.56	1.50
Highest ranked composition	32	5	5-	10	7.78	1.60

Examination of the data reveals that there are significant differences between raters in terms of the overall grade awarded; the least difference being of 5 points on a 10 point scale for ratings across the same compositions. Although there is more variability between raters in the highest ranked composition (SD = 1.60), the mid-ranked composition has a wider range of scores (2-8).

4.2. Influence of different composition elements on ratings

Recall that raters were asked to choose the best and worst feature of each composition from among *content, organization, grammar, register, mechanics* and *presentation*. If no judgement was given on a particular category, it was entered in the *no judgement* category. This category can be as revealing as positive and negative judgements, for if raters do not pay attention to some categories, this could have a significant effect on the final score (Gamaroff 2000).

The results for these analyses begin in Fig.1a, in which the best features of the lowest ranked composition identified by raters are shown. As can be seen from this data, 31.3% of all raters gave positive comments on *grammar* while *vocabulary* (21.9%) and *organization* (12.5%) were rated second and third respectively in order of importance. *Content*, on the other hand, was practically unattended (6.3%). The *no judgement* category registered 28.1% of the responses and it was assumed that raters did not find any feature of the text that deserved a specific mention.

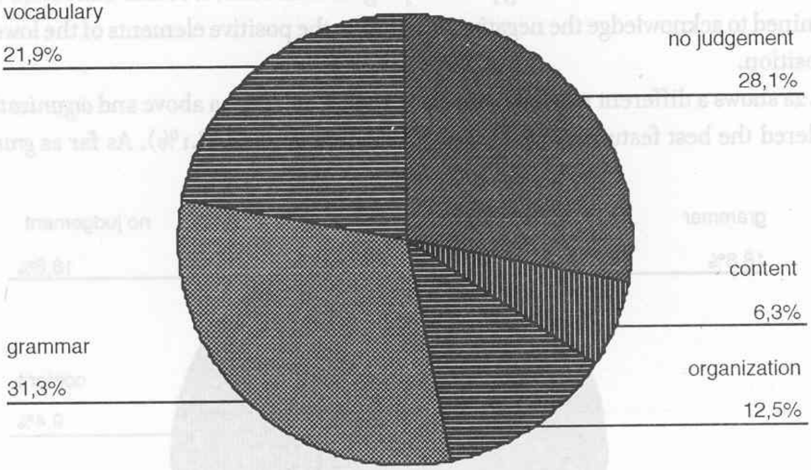

Figure 1a. Positive comments on the lowest ranked composition

In Fig.1b, the feature most often identified as the worst by all raters was *grammar*, which exhibited marked differences with the rest of the categories (87.5%) indicating an overall agreement among raters as to the importance attached to this category as a negative feature. *Content* and *organization* were the least often associated with the worst feature (3.1 % each). The

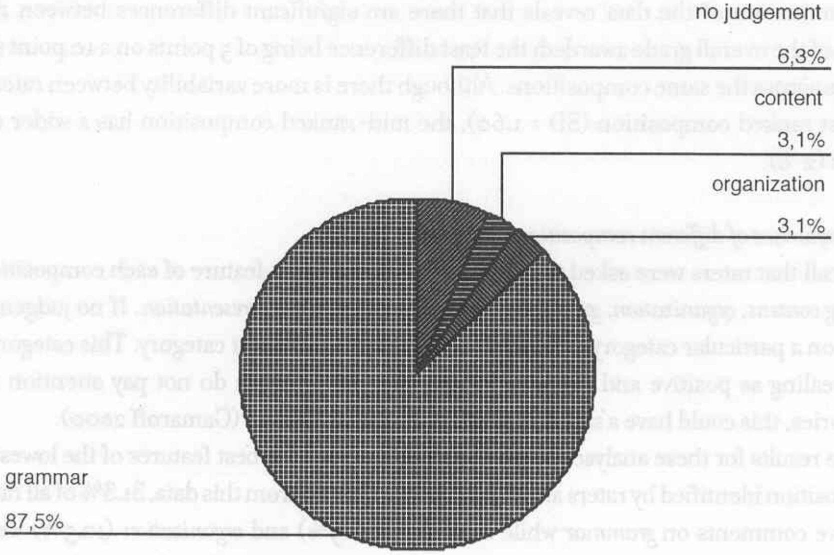

Figure 1b. Negative comments on the lowest ranked composition

no judgement category was chosen by 6.3% of the raters, a substantial difference as compared to the one recorded when considering positive judgements. Thus, it seems that raters were more determined to acknowledge the negative rather than the positive elements of the lowest ranked composition.

Fig.2a shows a different pattern from that produced in Fig.1a above and *organization* is now considered the best feature of the mid-ranked composition (53.1%). As far as *grammar* and

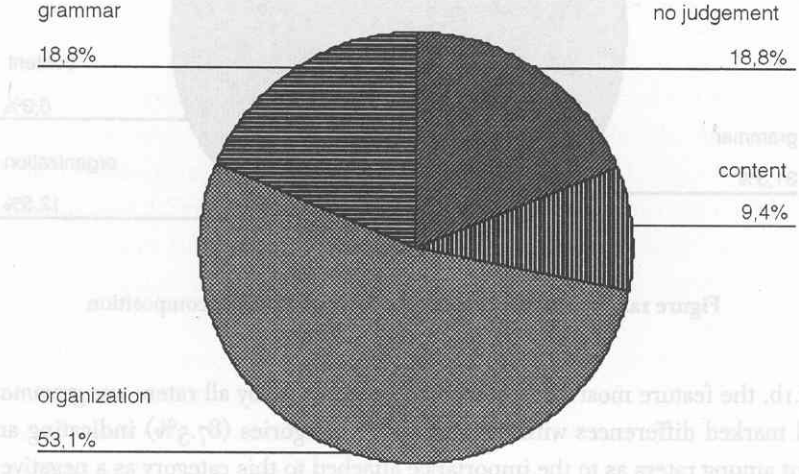

Figure 2a. Positive comments on the mid-ranked composition

content are concerned, 18.8% of the raters were positive about *grammar* and just 9.4% were positive about *content*. As can be seen, the *no judgement* category was given equal weight than *grammar* (18.8%).

In Fig.2b, it appears that *grammar* is again identified as the worst feature of the composition (34.4%). Raters also attended to *vocabulary* (18.8%) and this time greater emphasis was placed on *organization* and *content* (15.6% and 9.4%, respectively). *Presentation* was also listed (6.3%) and the *no judgement* category registered 15.6% of raters' responses. The latter fact seems to suggest that raters adopted a more balanced approach regarding the acknowledgement of positive and negative features for the mid-ranked composition.

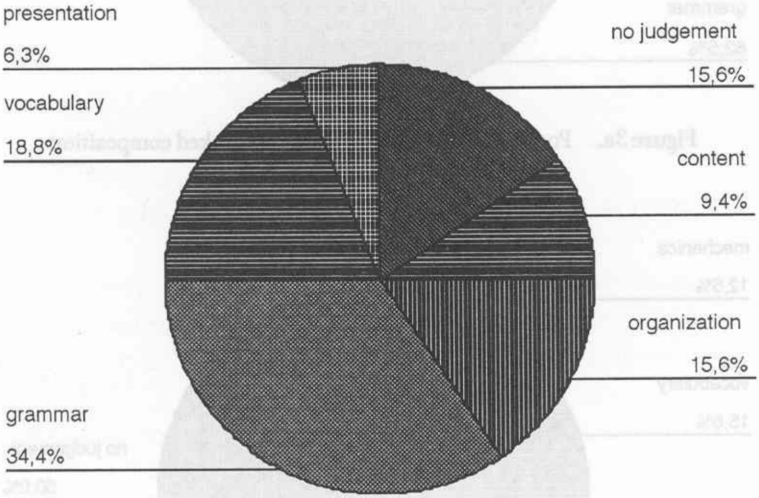

Figure 2b. Negative comments on the mid-ranked composition

Fig.3a displays the best features of the highest scored composition reported by raters. Examination of this table shows clearly rater agreement on *grammar* which is associated with the best feature (62.5%). On the other hand, *vocabulary* and *organization* only evidenced 9.4% of positive comments. Interestingly, a considerable number of raters (18.8%) gave no positive judgement to the highest ranked composition. Thus, it seems that raters are reluctant to acknowledge students' merits and give positive comments to compositions despite their being *good* compositions.

Finally, as Fig.3b shows, the feature of the highest ranked composition identified as the worst was *vocabulary* (15.6%). It is also noteworthy the fact that *grammar* and *mechanics* were given equal weight, since 12.5 % of the raters were negative on both categories. *Content* did not show marked differences from previous compositions and it was paid quite moderate attention by raters (6.3%). *Organization* was also hardly attributed any weight as a negative feature

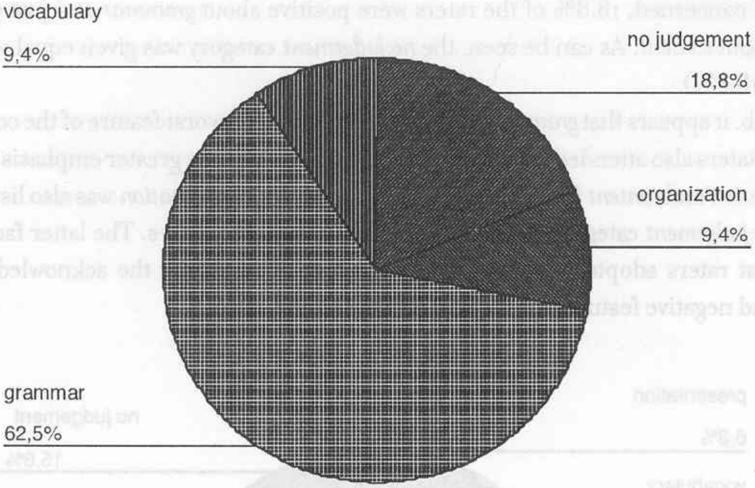

Figure 3a. Positive comments on the highest ranked composition

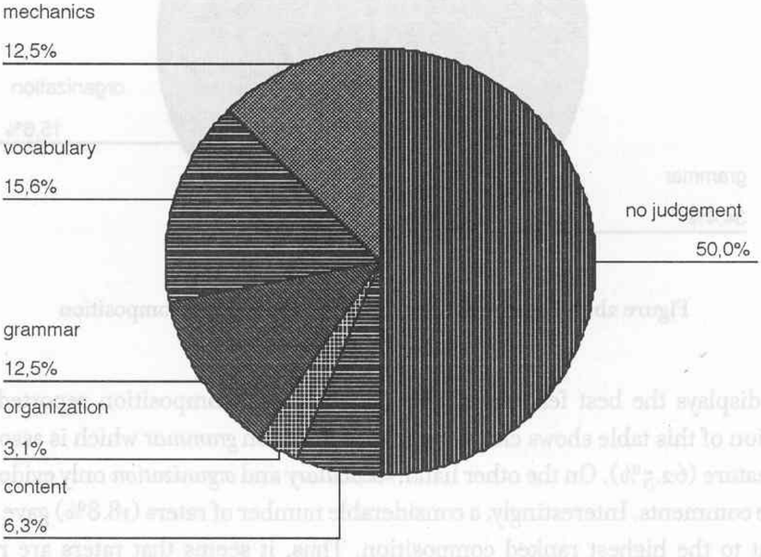

Figure 3b. Negative comments on the highest ranked composition

(3.1%). It seems that 50% of the raters considered that the errors or negative features were not serious enough to be mentioned for the highest ranked composition.

In short, with respect to the best features, raters seem to attend to *grammar* as a primary positive feature. Yet, raters appear to consider *organization* more important for middle scores.

In terms of negative features, *grammar* shows the same pattern, that is, it is identified as a primary negative feature. *Vocabulary* is, however, applied as the worst feature for the highest scored composition (see Santos 1988; Milanovic et al. 1996). *Content* is considered a positive feature across compositions although is hardly given any weight. It is interesting to note that *mechanics* is found to be an important negative feature for the highest scored composition. It seems likely that mechanical errors become more evident in high-ranked compositions because they are more salient.

Thus, the data suggest that raters vary their comments and reactions according to the perceived level of proficiency of the writer. Furthermore, particularly salient features of compositions seem to exert an important influence on raters' judgements.

4.3. *Error analysis across sentences*

The results yielded by the best and worst feature analyses parallel those of error analysis presented in isolated sentences. A list of fourteen original sentences, each containing one of seven error types related to the categories of *content*, *organization*, *grammar*, *vocabulary*, *register*, *mechanics* and *presentation* was distributed to all raters who were instructed to rank the errors on a 5-point level of acceptability (5 = very important). The mean of each error is shown in Table 3.

Table 3. Error analysis across raters

Error analysis	Category	Mean
1. I want to go to London to see the Statue of Liberty	Content	2.66
2. Thank you, I will accept your invitacion	Mechanics	2.44
3. If I went to London I would bought a lot of things	Grammar	4.25
4. I like speaking english very much	Mechanics	2.91
5. I find London very interesting but the food is wrong	Vocabulary	3.42
6. I would visit all the museums, furthermore, Big Ben	Organization	3.38
7. I have been to London and I have visited a few shops. Perhaps you think I'm stupid but don't worry about that	Register	2.94
8. One of my main hopeness is to visit London	Vocabulary	3.44
9. I like London. However, my parents and I will visit it next year	Organization	3.59
10. I don't like London and I think it's a horrible island	Content	3.34
11. A: Are you sure it would not be a problem for your cousin to put me up? B: No, silly	Register	3.06
12. I would visit the most importants places in London	Grammar	4.13
13. Sample of a student's handwriting	Presentation	3.61
14. Sample of a student's handwriting	Presentation	3.32

As we can observe, the means of error types connected to *grammar* in sentences 3 and 12 are the highest across sentences ($\overline{X}$ = 4.25 and $\overline{X}$ = 4.13 respectively). As far as *vocabulary* and *organization* is concerned, the means for error types related to *organization* in sentences 6 and 9 are $\overline{X}$ = 3.38 and $\overline{X}$ = 3.59 respectively, which are very similar to those provided by *vocabulary* in sentences 5 and 8 ($\overline{X}$ = 3.42 and $\overline{X}$ = 3.44 respectively). It is worth noting that the means for *presentation* errors contained in sentences 13 ($\overline{X}$ = 3.61) and 14 ($\overline{X}$ = 3.32) were higher than the means for *content* contained in sentences 1 ($\overline{X}$ = 2.66) and 10 ($\overline{X}$ = 3.34).

Thus, in spite of the basic differences in methodology between error gravity analysis based on isolated sentences and errors presented in extended discourse, similar results as to what kind of errors are considered to be most serious appear to have been reached. That is, raters maintain their greater concern with *grammar* followed by *vocabulary* and *organization* in order of importance.

4.4. The relationship between scores and judgements

The relationship between individual scores and judgements was also considered and a search for trends was made by means of the compositions. The data collected by the lowest ranked composition appear in Tables 4a and 4b of the Appendix. It can be seen from this data that individual raters categorised the different errors following their own criteria. The research shows that raters, do not agree on the classification of errors and that judgements on these issues might vary significantly across raters, since it becomes a matter of individual interpretation.

However, since it was the raters' perception and characterization of the students' writing that was important, the analysis of the relationship between scores and judgements is also quite revealing. If we observe the positive comments related to the lowest ranked composition (Table 4a), we realise that the two main categories emphasised by raters were *grammar* and *vocabulary*. However, it is interesting to note that both categories are mainly associated with the same sentence or part of it: "I think that the best soccer is played in England", which seems to have been made an impression on raters and is, accordingly, mentioned by eleven of the twenty-three raters who provided a positive comment on the composition.

For the negative comments on the lowest ranked composition (Table 4b), the majority of raters associated *grammar* with the misuse of the second conditional in English, a point which was clearly stressed by twenty-two of the thirty raters who made negative comments on the composition.

Therefore, if we consider the relationship between scores and judgements for the lowest ranked composition, we can see that despite the fact that similar positive and negative judgements were applied by raters they resulted in different scores. Thus, a score of 4 for one rater represented "I would + participle instead of I would + infinitive" (R4, Table 4b) and for another rater a score of 1 represented the same judgement: "Grammar errors, mostly conditional tenses" (R5, Table 4b).

Table 5a (see Appendix) shows the positive comments made about the mid-ranked composition. The feature most often identified as best by raters was *organization*. The analysis of individual judgements shows that raters were mainly influenced by the sentence found in the introduction of the composition: "A holiday in London would be fantastic", which was mentioned by seven of the nineteen raters who identified *organization* as the best feature of the composition.

As Table 5b (Appendix) shows, *grammar* was the feature identified as worst for the mid-ranked composition. *Vocabulary* was rated second in order of importance. However, it is interesting to note that once again both categories point to the same sentence: "I thing there is impressionant", which appears to be salient to most of the raters despite the fact they stressed different elements such as *grammar*, applied by eight of the twelve raters who chose this category or *vocabulary*, applied by the total number of raters who identified the latter category.

Furthermore, when we look at the relationship between scores and judgements for the mid-ranked composition, we observe again that the similarity of comments as far as the identification of positive and negative comments were concerned did not necessarily mean similar scores. Thus, a score of 7 for R1 (Table 5a) represented the same judgement as a score of 3 for R8 (Table 5a), that is: "Good introduction: A holiday in London would be fantastic".

Table 6a (Appendix) shows grammar as the best feature identified by raters (twenty raters on the whole) for the highest ranked composition. The most frequent comments made by raters seem to point to the sophistication of syntax and the use of verbal tenses (five raters) although some raters simply listed the category *grammar* as a global category (three raters) and some others emphasised the use of adverbs in the sentence: "I have always wanted" (two raters).

Finally, and despite the fact that 50% of the raters did not provide negative judgements for the highest ranked composition, as Table 6b (Appendix) shows, the most salient negative element was *vocabulary*, which was thought to be easy and repetitive. The most interesting point here is the inclusion of *mechanics* as a negative feature, which was ranked second together with *grammar* in order of seriousness. It appears that the misspelling of the word *heighs* was salient to many raters, who penalised the strongest students for an error generally considered a relatively minor error.

When we consider the relationship between scores and judgements for the highest ranked composition, we also observe that different scores between raters do not necessarily mean different judgements. Furthermore, raters may give equivalent scores but this does not mean that these scores represent the same. Thus, a score of 9 (Table 6a) for R1 represented: "Grammar: I have always wanted", but for R20 the same score represented: "Good use of connectors and good organization". Apparently, raters seem to follow different processes in arriving at these similar ratings. Thus, the agreement between raters on the ratings that have been awarded does not guarantee that raters have come to an agreement on what it is that they are rating.

5. Conclusions

The comparison of the holistic scores assigned by raters to the compositions revealed substantial variability in scores across raters. This is because *subjectivity* in rating behaviour is very difficult to control and even despite similar training in evaluation techniques, which is not the case in most usual education situations like the University Entrance Examination, raters may react differently to various aspects of the compositions. Thus, one rater will give high marks to a particular composition while another will consider the same composition to be weak or irritating and will label it as a low quality composition.

The proficiency level of the compositions might also influence raters who seem to adjust their marking behaviour according to the level of the composition. It has also been found that raters may differ in how to categorise errors and in the importance they give to the evaluative criteria (see section 4.2). Furthermore, raters may have similar judgements on different aspects of the compositions but this does not mean similar scores, as has been shown in section 4.4 of this study.

As far as the composition elements and error analyses are concerned, the results are very similar and they indicate that raters have the tendency to primarily point out negative rather than positive features. Raters put a great emphasis on *grammar*, which is found to be a critical factor in students' compositions as both a positive and a negative feature.

While raters seemed to have also focused on *vocabulary* and *organization*, it is cause for concern to note that the quality of *content* had no observable effect on the compositions' holistic scores despite the enormous impact of communicative approaches on the composing processes of students (Kroll, ed. 1990; Amengual and Herrera 2000). One realisation that arises from such results is that it is necessary to include both of these perspectives in the assessment of ESL written work so as to attend to the communicative demands as well as the linguistic demands of our academic communities.

Finally, and on the basis of these modest but relevant results, it is evident that raters who holistically score ESL compositions should be instructed to approach them differently in order not to be unduly influenced by insignificant errors or characteristics in the compositions that are easy to pick out but are irrelevant to effectiveness of communication. Rater training procedures may be, perhaps, the most direct way to modify grading practices and redress this situation (Sweedler-Brown 1993). Also, it is believed that discussions and moderation workshops will help to neutralise the problems of raters' subjective judgements so as to meet the linguistic and communicative needs of ESL students.

It is hoped that the results of the present study will contribute to our understanding of how raters make decisions about compositions and might provide some empirical basis for further research which would be of value in providing better training for raters.

Appendix

Table 4a. Scores and positive judgements for the lowest ranked composition

Raters	Score	Category	Raters' comment
R1	3	Grammar	'I think that the best soccer is played in England'
R2	3	——	——
R3	2	——	——
R4	4	Grammar	'I think that the best soccer is played in England'
R5	1	——	——
R6	2	Vocabulary	'soccer'
R7	1	——	——
R8	2	Vocabulary	'The best soccer in the world'
R9	2	——	——
R10	3	Content	Content
R11	1	Vocabulary	'A very famous shop center'
R12	2	Vocabulary	'The best soccer in the world'
R13	2	Organization	Organization: linearity although it may well be due to the lack of content
R14	2	Grammar	S/he tries to use conditional tenses although unsuccessfully
R15	1	——	——
R16	2	Vocabulary	'The best soccer in the world is played in England'
R17	3	Grammar	Short sentences
R18	3	Grammar	'I think that the best soccer in the world is played in England'
R19	3	Vocabulary	'soccer'
R20	6	Organization	'The last place I would visit in London' as conclusion
R21	4	——	
R22	1	Content	S/he Knows that London exists
R23	2	——	
R24	1	Grammar	'If I went'
R25	3	Organization	Organization of ideas
R26	2	Grammar	If I went to London
R27	4	Grammar	Variety: 'went', 'visit', 'see', 'I would like', 'take'....
R28	2	——	
R29	3	Organization	Communicative structure
R30	2	Vocabulary	'The best soccer in the world'
R31	3	Grammar	'The best soccer in the world is played in England'
R32	4	Grammar	'I think that the best soccer in the world is played in England'

Table 4b. Scores and negative judgements for the lowest ranked composition

Raters	Score	Errror Category	Raters' comment
R1	3	Grammar	'I would like visited': conditional structure
R2	3	Grammar	Misuse of conditional tenses
R3	2	Grammar	'I would went visited the real palace'
R4	4	Grammar	'I would + participle' instead of 'I would + infinitive'
R5	1	Grammar	Grammar errors, mostly conditional tenses
R6	2	Grammar	Misuse of conditional tenses
R7	1	Grammar	Misuse of Conditional tenses
R8	2	Grammar	'I would went in London would been'
R9	2	Grammar	Misuse of 2nd Conditional: 'would + simple past' / 'would + past participle'
R10	3	Grammar	Grammar
R11	1	Grammar	Verbs: 'would liked visited...'
R12	2	Grammar	'I would liked / visited / seen / bought / went / taked'
R13	2	Organization	Monotonous structure: 'I + verb' as introduction to new topics
R14	2	Grammar	'would + past participle'
R15	1	Content	Off topic: S/he doesn't answer Why s/he would visit certain places in London
R16	2	Grammar	Misuse of conditional tenses: 'I would toked photo with'
R17	3	Grammar	Misuse of Verbal tenses
R18	3	Grammar	Auxiliaries and Past Participle
R19	3	Grammar	Verb errors and misuse of conditional tenses
R20	6	Grammar	Verb errors
R21	4	Grammar	'I would liked visited'
R22	1	Grammar	Misuse of verb tenses
R23	2	--	—
R24	1	Grammar	Misuse of verb tenses
R25	3	Grammar	Morfosintaxis
R26	2	Grammar	'I would liked visited'
R27	4	Grammar	Modal + infinitive without 'to'
R28	2	---	—
R29	3	Grammar	Grammar correction
R30	2	Grammar	Morfosintaxis and verb tenses
R31	3	Grammar	Conditionals: 'I would went visited', 'I would taked photo'
R32	4	Grammar	'If I went of holiday in London I would liked visited in once tsh time the Big Ben'

Table 5a. Scores and positive judgements for the mid-ranked composition

Raters	Score	Category	Raters' comment
R1	7	Organization	Introduction: 'A holiday in London would be fantastic'
R2	5	—	—
R3	6	Organization	'Apart from this tower we can visit a lot of monuments...'
R4	6	Organization	Structure: 'The first thin we can say is that...'
R5	3	Organization	Coherence in development: opening, building and conclusion
R6	5	—	—
R7	5	Content	S/he tries to be original
R8	3	Organization	'1st sentence: 'A holiday in London would be fantastic'
R9	4	Organization	'The first thing we can say is that'
R10	5	Grammar	Grammar
R11	5	Organization	Introduction: 'A holiday in London would be fantastic'
R12	4	Organization	'The first thing we can say is that..' / 'In conclusion we can say that....'
R13	2	Organization	S/he tries to be coherent although at the end the organization is poor and incoherent
R14	4	Organization	S/he tries to give us a conclusion although it is not a very good one
R15	4	—	—
R16	5	Organization	Organization: 'apart from that', 'in conclusion'
R17	4	Grammar	Brief and short sentences
R18	6	Organization	1st paragraph and the expression 'apart from'
R19	5	Grammar	Verbal tense in 'a holiday in London would be fantastic'
R20	4	Organization	Introduction: 'A holiday in London would be fantastic'
R21	7	Organization	Structure: 'The first thing we can say is that'
R22	5	Content	Content original and involving personal experience
R23	3	—	—
R24	3	Grammar	Grammar structure: 'But the most important thing...'
R25	3	Organization	Introduction: quite good
R26	6	Organization	'A holiday in London would be fantastic' as a way of introduction
R27	8	Content	Variety and development of ideas
R28	2	—	—
R29	5	Organization	Good organization in the 1st part
R30	2	—	—
R31	6	Organization	'Apart from that tower we can visit'
R32	4	Organization	Introduction: 'A holiday in London would be fantastic'

Table 5b. Scores and negative judgements for the mid-ranked composition

Raters	Score	Errror Category	Raters' comment
R1	7	—	—
R2	5	—	—
R3	6	Vocabulary	'I thing there is impresionant'
R4	6	Vocabulary	Poor vocabulary: 'I thinG there is IMPRESIONANT'
R5	3	Grammar	Serious Grammar errors: Subject missing, 'I thing...'
R6	5	Content	Change of topic; from people to monuments without guiding us
R7	5	Grammar	'I thing there is impresionant'
R8	3	Organization	'In conclusion we can say that London is a strange place'
R9	4	Grammar	Subject missing
R10	5	Presentation	Presentation
R11	5	Vocabulary	'I thing there is impresionant'
R12	4	Grammar	Sentences without a subject; 'I thing there is impressionant'
R13	2	Grammar	'choppy style' and misuse of punctuation. Abuse of structures like ('We can + Verb) taken from Spanish
R14	4	Content	Contradiction: S/he criticises racism in London and then s/he talks about 'illicit' people which shows S/he a racist. Change of topic from 'places of interest' to 'people'
R15	4	Organization	Confusing, poor organization: 'We can find black...and from other country have visited' (it is not coherent)
R16	5	Grammar	Grammar: 'I thing there is impresionant'; relative clause and change of subject from 'we' to 'I'
R17	4	Organization	Conclusion
R18	6	Vocabulary	'impresionant', 'illicit'
R19	5	Organization	Lack of cohesion: poor structure
R20	4	Grammar	Relative pronoun is missing; it impedes comprehension: '...people from India and from other country (WHO?) have visited London for one time'
R21	7	Grammar	'(Subject missing) could observe that...'
R22	5	Grammar	Grammar errors in the second part
R23	3	—	—
R24	3	Grammar	'I thing there is impresionant'; 'Could observe that'
R25	3	Organization	Horrible ending
R26	6	Vocabulary	'I thing there is impresionant'
R27	8	Presentation	Full of crossings out
R28	2	—	—
R29	5	Content	Off topic
R30	2	—	—
R31	6	Grammar	Grammar: 'I thing there is impresionant'
R32	4	Grammar	Grammar in the sentence: 'I thing there is impressionant'

Table 6a. Scores and positive judgements for the highest ranked composition

Raters	Score	Category	Raters' comment
R1	9	Grammar	'I have always wanted'; 'All of them'
R2	8	—	—
R3	10	Grammar	'I would probably contract a guide in order to visit the mos important...'
R4	9	Grammar	Use of subordinate clauses
R5	6	Grammar	Use of verbal tenses
R6	9	Grammar	'In order to'
R7	5	—	—
R8	5	Grammar	'I have always wanted'
R9	8	Grammar	Use of conjunctions and sophisticated syntax
R10	6	Grammar	Grammar
R11	5	Grammar	'That's the reason why'
R12	6	Vocabulary	Vocabulary: 'I hope that some day my wish will come true'
R13	5	Grammar	Verbal tenses and sentence structure and comparatives
R14	7	Grammar	Good grammar makes it fluent
R15	8	—	—
R16	9	Organization	Structure: Introduction, body and conclusion
R17	7	Grammar	Grammar
R18	10	Grammar	Good position of adverbs
R19	8	Grammar	'That's the reason why'
R20	9	Organization	Good use of connectors; good organization and structure
R21	9	Grammar	'I don't know all the places'
R22	8	Grammar	Complex sentence structure
R23	5	—	—
R24	9	Grammar	'I hope that some day my wish will come true'
R25	8	—	—
R26	8	Grammar	'I have never seen a river so that's the reason why I would like to visit the Thames River'
R27	10	Organization	Organization: introduction, body and conclusion
R28	8	—	—
R29	8	Grammar	Variety of grammatical structures
R30	9	Vocabulary	Vocabulary expressions: 'My wish will come true'
R31	9	Grammar	'From my point of view this visit is more boring than...'
R32	9	Vocabulary	'I hope that some day my wish will come true'

Table 6b. Scores and negative judgements for the highest ranked composition

Raters	Score	Errror Category	Raters' comment
R1	9	—	—
R2	8	—	—
R3	10	—	—
R4	9	—	—
R5	6	Vocabulary	Vocabulary: easy and repetitive 'to go', 'I like'
R6	9	Grammar	Subject missing 'it': 'I think that is'
R7	5	Content	Simplistic content
R8	5	—	—
R9	8	Vocabulary	Poor vocabulary: repetitive (e.g. 'visit')
R10	6	—	—
R11	5	—	—
R12	6	Content	Content weak and not well argued
R13	5	Mechanics	Punctuation (before 'so' or 'from my point of view' or 'but'. Minor errors 'heighs' instead of 'heights' but they are not significant
R14	7	Organization	Last paragraph it's too short
R15	8	—	—
R16	9	—	—
R17	7	—	—
R18	10	Mechanics	Spelling: 'heighs'
R19	8	Grammar	'Another place that I would probably go'
R20	9	—	—
R21	9	—	—
R22	8	Vocabulary	Repetitive vocabulary
R23	5	—	—
R24	9	Grammar	'I think that IS a very beautiful...'
R25	8	Vocabulary	Poor vocabulary and repetitive structures (e.g. 'visit')
R26	8	Mechanics	'I like heighs very much'
R27	10	—	—
R28	8	—	—
R29	8	Grammar	Minor grammatical errors
R30	9	—	—
R31	9	Vocabulary	'I would probably contract a guide'
R32	9	Mechanics	'...because I like heighs very much'

6. REFERENCES

AMENGUAL, M. and HERRERA, H.

2000 *Raters Assumptions about Form and Content.* XVIII AESLA Conference. Barcelona.

DAVIES, E.E.

1983 Error evaluation: the importance of viewpoint. *ELT Journal* 37, 304-311.

DORDICK, M.

1996 Testing for a hierarchy of the communicative interference value of ESL errors. *System,* Vol. 24, N° 3. 299-308.

FREEDMAN, S.

1979 How characteristics of student compositions influence teachers' evaluations. *Journal of Educational Psychology*, 71, 328-338.

GAMAROFF, R.

2000 Rater reliability in language assessment: the bug of all bears. *System* 28, 31-53.

HAMP-LYONS, L.

1991 Scoring procedures for ESL contexts. In Hamp-Lyons (ed), 241-246.

HAMP-LYONS, L. (ed.)

1991 *Assessing Second Language Writing in Academic Contexts*. Norwood, N.J.: *Ablex Publishing Corporation*.

HERRERA SOLER, H.

1999 Is the English test in the spanish university entrance examination as discriminating as it should be? *Estudios Ingleses de la Universidad Complutense*, No. 7, 89-109.

HOMBURG, T.J.

1984 Holistic evaluation of ESL compositions: can it be validated objectively? *TESOL Quartely*, 18, 27-45.

HUOT, B.

1990a The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research*, 60, 2: 237-263.

KROLL, B. (ed.)

1990 Second language writing. *Research Insights for the Classroom.* Cambridge: Cambridge University Press.

MCDANIEL, B.A.

1985 Ratings vs. equity in the evaluation of writing. Paper presented at the 36th Annual Conference on College Composition and Communication, Minneapolis, MN.

MILANOVIC, M. and SAVILLE, N.,eds.

1996 *Performance. Testing, Cognition and Assessment*. Cambridge: Cambridge University Press.

MILANOVIC, M.; SAVILLE, N. and SHUHONG, S.

1996 A study of the decision-making behaviour of composition markers. In Milanovic and Saville, eds., 92-114.

MULLEN, K.

1980 Evaluating writing proficiency in ESL. In Oller Jr. and Perkins, eds., 160-170.

OLLER Jr.; J.W.

1979 *Language Tests at School.* London: Longman

OLLER, Jr.; J.W. and & K. PERKINS, eds.

1980 *Research in Language Testing.* Rowel, MA: Newbury House.

SANTOS, T.

1988 Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quartely.* Vol. 22 (4), 69-90.

SWEEDLER-BROWN, C.O.

1993 ESL composition evaluation: the influence of sentence-level and rhetorical features. *Journal of Second Language Writing* 2 (1), 3-17.

VAN, R.J.; MEYER, D.E. & LORENZ, F.O.

1991 Error gravity: faculty response to errrors in the written discourse of nonnative speakers of English. In Hamp-Lyons (ed.), 181-195.

VAUGHAN, C.

1991 Holistic assessment: What goes on in the rater's mind? In Hamp-Lyons (ed.), 111-125.