RESLA, 14 (2000-2001), 161-179

THE EFFECT OF GENDER AND WORKING PLACE OF RATERS ON UNIVERSITY ENTRANCE EXAMINATION SCORES*

HONESTO HERRERA SOLER
Universidad Complutense de Madrid

ABSTRACT. *Not only validity but also reliability must be considered as prototypical features in any test. When dealing with objective tests most researchers have centred their attention on validity whereas reliability has been associated with subjective tests. Ours will be a two-phase approach to study the effect of raters' gender and working place on reliability. In the first phase the first 30 ratings of 15 examiners assessing exams of the Complutense University Entrance Examination (CUEE) are analysed, and in the second we study the ratings of 20 examiners judging a sample of five exams of the CUEE. Significant differences are found in terms of gender and working place of raters, though these results can be relevant they must be taken with some caution because of the nature of the samples.*

KEYWORDS. *University Entrance Examination, rater, inter-raters, gender, working place, assessment, analytic, holistic, content, form.*

RESUMEN. *No sólo la validez sino también la fiabilidad debe considerarse un rasgo prototípico de cualquier test. En los tests objetivos el interés de los investigadores suele centrarse en la validez mientras que la fiabilidad se asocia con las pruebas de carácter subjetivo. En este estudio se analiza el efecto del género y del lugar de trabajo del corrector en la fiabilidad en dos fases. En la primera fase se valoran las calificaciones de los primeros 30 exámenes de 15 profesores que corrigen las pruebas de inglés de la selectividad. En la segunda fase se analizan las calificaciones de 20 correctores que evalúan una muestra de 5 exámenes. Diferencias significativas se dan en función del género y del lugar de trabajo de los correctores, aunque los resultados deben tomarse con cierta cautela debido a la naturaleza de las muestras.*

PALABRAS CLAVE. *Selectividad, corrector, correctores distintos, género, lugar de trabajo, evaluación, analítico, holístico, contenido, forma.*

1. INTRODUCTION

The blossoming of language testing research that we have witnessed in this period has provided us with a rich variety of research approaches and tools, while at the same time broadening the research questions that are being investigated. While

> *this research has deepened our understanding of the factors and processes that affect* performance *on language tests, it has also revealed lacunae in our knowledge and pointed to new areas for research (Bachman 2000: 2).*

1.1. *Gender and working place of raters*

This study focuses on the working place and gender of raters, two factors, that would belong to these lacunae which demand further research. So far, most research questions on raters have been about their criteria, rating scales or personal judgements, especially in subjective tests, but little has been done on the influence on scoring of the raters' working place and gender.

It is assumed in the University Entrance Examination objective subtests that, at least in theory, there should not be room for different ratings, though, unfortunately, the marking scheme, a basic feature of testing, is not always clearly defined and a considerable amount of subjective components can be found in objective subtests (Herrera 1999). As for subjective tests there are no marking schemes, nor clear cut defined guidelines for answers, and, consequently, more fluctuations than agreements on the test-takers' answers can be found among raters.

Thus, a test can be judged as good or not so good according to the strictness and leniency of the raters, and substantial variability according to different criteria is revealed again and again. This has led us to consider the so-called subjective subtests of the Complutense University Entrance Examination (henceforth CUEE) as the main issue of this study. We do not know what is going on in the raters' mind when they are assessing an essay, but we can conclude from the results that their behaviour differs depending on whether the rater is male or female, or works at a Secondary School or at University.

1.2. *Criteria*

Though holistic assessment is a technique based on the assumption that if scoring rubrics are provided inter-rater reliability can be high (Brown et al. 1992; Shavelson et al. 1992), findings are not consistent with this theory and a wide range of scores among raters is not uncommon. The search for achieving an overall reliable score based on the judgement of specific criteria continues year after year and researchers go on working to build a standard rating scale that might be an answer to this issue. Meanwhile, they tackle the problem either by developing a specific rating scale for each situation (Carson et al. 1990) or by adapting some of the already existing versions of English L2 rating scales (Pennington and So 1993).

An approach to the latter tendency may be read in the set of descriptors for the subjective test given by the CUEE, where traditional rating scales are adapted. On the whole, there is not much definition and specification of the different components. The rationale of the rating guidelines provided is to emphasise the achievement of a reasonable communicative competence. On this basis, both comments on the structure,

content and vocabulary, and warnings of pasting of chunks from the reading are made; but little information is found on the weighting of the different elements.

The designers' instructions are so open that consistent judgement criteria are becoming a difficult task, even among expert raters. This fact leads raters to assess papers from their personal point of view. That means the judgement criteria of each rater will be affected by the linguistic theories he / she has on language usage, on the students' learning process and on his / her teaching ideas. This background will underlie the rating process, where raters will try to match what they perceive in the performance they have to grade with their ideas on what this performance should be.

1.3. *Rating conditions*

The anonymity of the CUEE tests makes things different as far as the influence of raters on grading is concerned. The raters go beyond the classroom boundaries, where they know each student and they have their own expectations in keeping with the teaching-learning interaction process throughout the year. Unidentified exams will allow us an in depth study of traditional fields of research on raters: experts versus non-experts, or the influence of raters' gender on scoring behaviour.

Aware as we are of the circumstances of the Entrance Examination we take advantage of the situation and concentrate our efforts on the possible effects of the raters' gender and working place on their rating behaviour. Taking for granted that most of the raters have some sort of experience in grading CUEE and that the only cue to identify an exam is a number on the top right hand side of the exam paper our aim is to delve into the inter-raters' reliability.

A thorough analysis of rating behaviour leads to the study of whether an analytic approach provides more information than a holistic assessment when dealing with the essay, where not only the number of categories but also their relative weighting is considered. The answer to the last issue has been different among researchers. Whereas Diederich et al. (1961) assigned 10 points to the first two traits (ideas, forms) and five to the other traits (flavour, mechanics and wording), Jacobs et al. (1981) proposed more weighting to the first two components (content and organisation). Hamp-Lyons (1991) synthesised the multiple-trait scoring to ideas, rhetorical features and language control. If these series of proposals are summed up the main components of a standard rating scale could be reduced to two prototypical features with the same weighting: content and form.

Following both holistic (that is, what the Examination Board demands) and analytic criteria, we will study if there are differences in rating in terms of gender and working place not only between open questions and compositions but also between content and form in a writing text. It is assumed that analytic scorings provide more information (Jacobs et al. 1981; Raimes 1990; Hamp-Lyons 1991) and are more reliable than holistic scorings (Huot 1990; Hamp-Lyons 1991, 1995), though the Entrance Examination design is holistically oriented. Thus, in this paper we will concentrate on the reliability of raters, first on the holistic parameters required by the

Entrance Examination Board, and then on the analytic parameters demanded in this study, as a source of error measurement in assessing candidates' performance. Our research question, therefore, will be whether scores are significantly affected by the gender and working place of raters or if the different groups of raters give the same or similar scores to the same subjective subtests.

2. RESEARCH

2.1. *Need for a standard scale for rating a writing text.*

A glance at the literature will enable us to see how some raters concentrate almost exclusively on a paper's bad points, while others tend to mention those aspects that benefit the examinees. The concern about a standard scale for rating a writing text is not new. In Cambridge Examinations, FCE (First Certificate in English) and CPE (Certificate of Proficiency in English), eleven composition elements are listed[1]: Most of these elements and others which have not been pointed out have been in the researchers' mind as they have worked out their own scoring schemes.

Diederich et al. (1961), pioneers in developing analytic rating scales, identified five major traits when rating the compositions of English L1: ideas, forms, flavour, mechanics and wording, and emphasised the relevance of the first two traits. Stewart and Grobe (1979) highlighted length and accuracy in the field of writing English L2. The next step, among researchers, was to stress the relevance either of content or form.

Freedman (1979) showed that content was the most significant factor in the judgement of an essay, while Grobe (1981) found that raters were primarily influenced by vocabulary diversity. Jacobs et al. (1981) went further and developed a five-component scale: content, organisation, vocabulary, language use and mechanics. They stressed the concept of communicative effectiveness and emphasised the first two components rather than mechanics, a component on which teachers tend to place most emphasis when grading students' compositions. A decade later, Hamp-Lyons (1991) created the "Michigan Writing Assessment Scoring Guide" with three components: ideas and arguments, rhetorical features and language control with detailed descriptions for each of six levels. More recently, Sasaki (1999) has developed an analytic rating scale for Japanese L1 writing where "appeal to the readers and social awareness" stand out as new elements of a rating scale highlighting different qualities of the composition than the traditional scales.

2.2. *Interraters' reliability*

2.2.1. Personal criteria

Raters may give equivalent scores but this does not mean that these scores represent the same, since raters may differ in the importance they attach to different criteria, or their rating can be different because of their way of interpreting the same criteria. It is the latter aspect that Vaugham (1991) and Russikoff (1994) have dealt with. Vaugham taking similar training as a point of departure noted that the assessment

of an *essay* could be different, whereas Rusikoff emphasised the influence of the rater's background and personal interpretation on the rating. He thoroughly examined the different aspects that can intervene in the inter-raters' reliability: raters' perception of their role, imprecise criteria for scoring and confusion between inaccurate and non-standard structures. Despite that, little is known about the raters' decision processes. There is still lack of information both on the rank of the components of a rating scale and on the weighting of the raters' priorities. The rater's criteria related to grammar, vocabulary, structure and all the other elements that affect comprehension and coherence are applied to each exam on the spur of the moment. This rating behaviour explains why raters' personal responses to a script vary so much. One rater will give low marks to a particular *essay* while another will consider the same composition to be original and will label it as a good quality essay. More recently, Bachman (1995) focuses her studies on grammar ratings.

2.2.2. Gender

Gender is another issue that can also affect the inter-raters' reliability. References are abundant and many studies have been carried out on gender as an element of information that can explain a student performance. These differences in gender may be innately, psychologically or socio-culturally determined, or may show a bias in the rater grading his/her students.

There are studies on female and male linguistic behaviour in Gass and Varonis (1986) and Pica et al. (1989), who report men dominance in conversation, a view, which is supported in Shehadeh's (1999) studies on cross-gender conversation analysis. In this study it is suggested that men take advantage of the conversation to promote their performance production ability, whereas women utilise their conversation to promote their comprehension ability. Berninger et al. (1996) find boys to be lower in compositional fluency than girls but not in compositional quality. This line of research is also followed by Celce-Murcia (1997). She argues that male-female interactional differences have socio-cultural as well as gender-based origins. These findings raise the question of whether gender differences may have implications not only for L2 learning among students but also for assessment among raters.

Both pupils' and raters' gender may have a significant effect on rating. Hamp-Lyons (1990) points out that if the rater has some information of the student it is possible that sex, experiential background, profession, amount of exposure to L2 writing, etc. may affect the raters' behaviour. Both male and female raters give lower marks to the same piece of science writing when assigned to girls than when assigned to boys (Goddard-Spear 1983), the same article is rated more highly when it is attributed to John McKay than to Joan T. McKay. Rating changes when the subject is beyond the boundaries of the scientific field, where boys have traditionally been considered brighter than girls. Gipps and Murphy (1994) give evidence that in design and technology surveys female raters grade girls' work higher than boys'.

Arguments for and against are described with respect to raters' gender. Whereas for Gyagenda and Engelhard (1998) the gender effect is significant, for Baird (1998), there is no consistent evidence of gender bias in rating.

3. METHOD

3.1. *Conditions*

The constraints of the English test performance of June 1999 of the CUEE such as: control of students, identification, administration of the test, invigilation, etc., are the same as for other subjects like Mathematics, Philosophy, History or Spanish Language. It is a University Entrance Exam where the Administration asks the raters to discriminate as much as possible to rank students according to their level.

In this study our attention is focused on reliability and we investigate the gender and the working place of the rater, two specific features within the scope of inter-rater reliability. Thus, it is necessary to describe the raters' characteristics in relation to these two variables. They are male or female, working either at the University or in Secondary Education. Women raters outnumbers men raters, which is normal since there are fewer men than women in L2 teaching. This makes it difficult to find the same amount of raters for each category. Thus we have been obliged to study raters' reliability in two phases. In the first, the aim is to see if in a large sample, in which each rater grades his / her own tests, evidence of inter-rater differences due to gender and working place are found, whereas in the second our concern is the rating of the same tests by all the raters.

Anonymity is maintained throughout the rating process. Raters do not know which educational institution students come from, nor their names. It is only once that they have handed over the marked tests that they are allowed to know the educational institution they have assessed.

3.2. *Subjects*

In the first phase 10 female raters - five of them working at University and 5 in Secondary Education and 5 male raters working at University take part. The target is the ratings of 150 different subjects assessed by each group of raters. They were told to note down the first 30 students they scored. That means 450 subjects and 15 different centres, a sample large enough to carry out any statistical test. All the raters accepted the invitation to participate in this study.

This research demands a second phase in which the grading of the same *essays* by all the raters is analysed. To the 15 raters of the first phase 5 more are added. They come from Secondary Education and have experience of rating CUEE

As for the subjects, radomness is taken for granted because each rater was given no more than 200 tests on no other criteria, from CUEE Board, than that of distributing a similar number of examinations to each evaluator. Furthermore, the Administration did not know which raters would provide the data for this study.

3.3. *Components of the English Test in the CUEE*

The English test (ET)[2] consists of five different subtests based on a reading passage. Two of them can be considered subjective subtests: open questions drawn from the reading and an essay on topics related to the reading passage, whereas the other three are classified as objective subtests: two True / False items, lexical comprehension, in which examinees suggest synonyms for four underlined words or phrases from the reading passage, and finally a grammatical subtest, in which examinees complete sentences by using the terms given in parentheses or just filling in the gaps (table 1). As we are concerned with the raters' performance rather than with the level of difficulty of the different subtests I will not describe the different details on which the test was made up.

This examination is intended to assess the students' proficiency level in relation to the others, for the purposes of choice of career studies being assigned on a basis of rating hierarchy. This design should be evaluated in terms of how well it serves the utilitarian purpose for which it has been constructed. If it helps to spread students out into a normal distribution so that scores between performances are very different, the ET will fulfil the aims for which the test has been drawn up. If, on the contrary, there are items which do not accomplish the utilitarian purpose the ET has been built for and the results do not spread the scores out as expected, the validity of these items will be in jeopardy. In addition to this possibility of questioning the validity of a test on a basis of sample distribution there is also the risk, which is the object of this research, that the utilitarian purpose the test was designed for will not be fulfilled because of the raters' performance in terms of their gender and working place.

3.4. *Scoring*

Looking for reliability, the administrators of the entrance examination give rating cues for every item. It is assumed that with similar criteria there will be little room for disparity in scoring among raters, especially in the objective subtest, as they are considered items of a paper and pencil test. The following scoring scheme is provided for the raters together with general instructions on marking vocabulary, structures, organisation or content of the text (Table 1).

Table 1

Item	Score	Competence	Type of the item	Technique
1.	0 - 2	communicative	Subjective	Open answer
2.	0 - 2	comprehension	Objective	True / False
3.	0 - 1	lexis	Objective	Matching
4.	0 - 2	syntax	Objective	Cloze
5.	0 - 3	communicative	Subjective	Non-directed essay

Scoring instructions

Raters are asked not to use more than two decimals in the final results they deliver to the Examination Board (better *1.5* than *1.55*). This suggestion leads raters either to use one decimal in scoring each item or to use two whenever they wish to be accurate and only to round scores at the last moment on summing up. From the outset the student knows the weighting of each item in the final score as this appears in brackets following each question.

3.5. *Procedure*

In all test designs efforts are made to be as close as possible to real life performance. In the CUEE we rely on the real scoring behaviour of the raters. It is not an artificial but a real situation. There is no pilot test and they do what they are supposed to do. This enables us, in the first stage of this study, to face a real rating situation and, as a result, the data obtained from the raters are the data they provide to the Examination Board.

In the second stage raters work under experimental conditions and they are asked to follow specific instructions on the rating of the same sample of exams.

Thus, the scoring behaviour of raters is analysed in two phases:

In the first, the following issues are studied:

1. The scoring of the open question items under the condition of gender and working place,
2. Average scores among the categories of raters between the subjective and objective subtests,
3. The rating of the open questions and the essay.

In the second stage, our interest is focused on:

1. The raters' scoring behaviour towards, on the one hand, the open questions and *essay* and, on the other, form and content within the essay subtest,
2. The raters' performance depending on the quality of the essay.

4. RESULTS AND DISCUSSION

4.1. *The scoring of the open question items under the condition of gender and working place*

In the first stage we begin by presenting tables 2 and 3, where an illustration on the female raters' performance in keeping with their working place is offered. Both groups of raters keep within the lower and upper limits, 0 and 2 respectively, of the rating scale proposed by the Administration Board, but their rating behaviour is quite different. Whereas the University group of raters resorts to quartiles .25, .50, .75, etc., some Secondary School raters go beyond quartiles and their level of discrimination reaches each multiple of five. A scale of nine intervals for University female raters becomes a scale of 26 intervals in Secondary School female raters, though most of the frequencies cluster around unit quartiles in both groups.

Table 2

Scale of values	Absolute Frequencies	Relative Frequencies	Accumulated Frequencies
,00	7	4,7	4,7
,25	19	12,7	17,3
,50	28	18,7	36,0
75	10	6,7	42,7
1,00	21	14,0	56,7
1,25	11	7,3	64,0
1,50	17	11,3	75,3
1,75	19	12,7	88,0
2,00	18	12,0	100,0
Total	150	100,0	

Open question subtest. University female raters

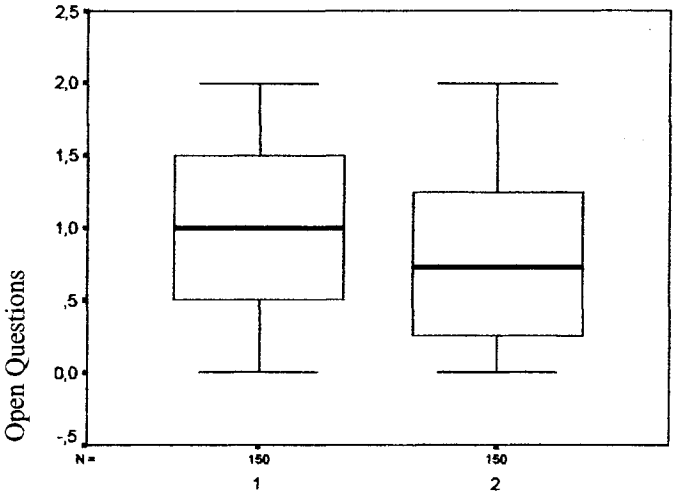
Table 3

Scale of values	Absolute Frequencies	Relative Frequencies	Accumulative Frequencies
,00	8	5,3	5,3
,05	1	,7	6,0
,10	2	1,3	7,3
,15	2	1,3	8,7
,20	7	4,7	13,3
,25	21	14,0	27,3
,30	4	2,7	30,0
,40	1	,7	30,7
,50	22	14,7	45,3
,55	1	,7	46,0
,60	1	,7	46,7
,70	5	3,3	50,0
,75	8	5,3	55,3
,80	1	,7	56,0
,85	1	,7	56,7
,90	2	1,3	58,0
1,00	14	9,3	67,3
1,10	5	3,3	70,7
1,25	9	6,0	76,7
1,35	1	,7	77,3
1,50	11	7,3	84,7
1,60	1	,7	85,3
1,70	1	,7	86,0
1,75	8	5,3	91,3
1,80	1	,7	92,0
2,00	12	8,0	100,0
Total	150	100,0	
Total	150	100,0	

Open question subtest. Secondary school female raters

These scoring differences could be explained on the basis of a global versus a detailed approach. On account of the daily teaching task of Secondary School raters with the test takers, their assessment could demand more accuracy and precision on each small lexical or grammatical detail. On the other hand, the teaching environment of University raters, more focused on specific issues of research fields, may encourage them to be more concerned with intelligibility than with errors. There is no reason for them to go beyond the unit quartiles. These raters fulfil to a lesser degree than the Secondary School raters the Examination Board expectations on discriminative scoring. There are different scoring procedures between these two groups, but we do not have information to choose one rather than the other.

Figure 1



Working Place 1. University 2. Secondary School

We do not know which rating behaviour is better, but a look at fig.1 gives a clearer idea of which category scores higher. It is not a matter of which is closer to the correct interpretation, but the Box Plot gives a first idea of strictness versus leniency according to the working place. It can be seen that the median and inter-quartiles of the Secondary School female raters is lower than that of the University female raters. The statistical test applied shows significant differences among the two female rater categories ($t = 2.818$, $p < .005$). Thus, it can be put forward that the working place is a factor among female raters which influences the reliability of the "open question items".

4.2. *Comparison between subjective and objective scoring behaviour*

The starting point for the analysis of the raters' behaviour in accordance with their working place and gender is Table 4. Several issues should be taken into account not only in the subjective part of the CUEE test, which theoretically could be expected, but also in the objective part of the test, which, in principle, should not be expected.

Men raters' leniency scoring behaviour in the objective subtests of the exam is evident while women raters' behaviour is similar whether they work at Secondary School or at University. Male raters' mean with values of 6.2618 is considerably higher than female raters' mean of 5.3368.

Table 4

Objective Subtests	Gender. 1 Male 2 Female	Working Place 1 University 2 Secondary School	Mean	Standard Deviation	Number
	1	1	6,2618	2,3867	150
		Total	6,2618	2,3867	150
	2	1	5,3439	2,4090	150
		2	5,3298	2,3133	150
		Total	5,3368	2,3577	300
	Total	1	5,8028	2,4376	300
		2	5,3298	2,3133	150
		Total	5,6451	2,4047	450
Subjective Subtests	1	1	6,1360	2,0834	150
		Total	6,1360	2,0834	150
	2	1	5,4533	2,5180	150
		2	3,8913	2,4819	150
		Total	4,6723	2,6156	300
	Total	1	5,7947	2,3322	300
		2	3,8913	2,4819	150
		Total	5,1602	2,5443	450

Descriptive statistics of objective and subjective subtests
** The raw scores have been transformed to allow comparisons.

In the subjective subtests university female raters change their scoring behaviour and their mean is closer to that of university male raters than to that of Secondary School female raters. In this occasion female raters are nearer to male raters´ leniency: with mean values of 5.4533 and 6.1360 respectively, whereas strictness appears to be a defining characteristic of Secondary School female raters as their mean is a mere 3.8913.

The ANOVA shows that there are significant differences between the different components of the test: objective and subjective subtests, $F_{1,447} = 32.333$ and $p<.001$, and interaction effect is observed between the components of the test and the working factor: $F_{1,447} = 38.253$ and $p<.001$. There are also two principle effects between groups related to gender and working place with $F_{1,447} = 10.816$ and $p<.001$ for the gender factor and $F_{1,447} = 10.488$ and $p< .001$ for the working place factor.

The reading of these data demands some comments on the components of each subcategory of tests. The objective subtest is thought out in the designers' mind as a

series of items where only one answer should be taken as correct for each item. But there is not always a coincidence or general agreement between what it is and what it should be. And that is what usually happens in the Entrance Examination, an exam which, to avoid leaks of information, has not been piloted. Thus, the instructions, the wording of the items or the context can lead to more than one item being acceptable. It is up to each rater either to interpret or admit an unexpected, but appropriate answer or to stick to the proposed Examination Board marking scheme. This fact explains the different scoring behaviour.

Again, it is necessary to look into the components of these objective subtests in order to find a reasonable explanation for these scoring differences. Then, examining each item of the components, reasons for broad and narrow interpretations are found. There is room for a wide range of interpretations not only in the True / False questions, where just the mere indication of the evidence required in each item can be taken as acceptable or not, but also in the grammatical subtests where sometimes several answers are acceptable. These results lead us to suggest further research on this issue to see if the differences persist, once we get rid of the extraneous variables through accurate and well-defined pen and pencil designs in the Entrance Examination. If there are still significant differences, then Celce-Murcia's (1997) research on gender-based origins should be taken into account.

The range between Secondary School and University averages in the subjective subtests seriously questions inter-rater reliability. These results demand specific research in order to find an explanation for this different scoring behaviour among raters. There is work to be done on the personal point of views of raters: (error interpretation: quantity versus quality, error hierarchy, levels of communicative effectiveness etc.) according to their working place. These or other criteria lead Secondary School raters to be stricter than University raters. Their judgements about a learner's level of skills and knowledge do not coincide with those of the University raters.

If we were to go deeper into this issue we would have to take into account that in educational assessment now, the same as in psychometric testing previously, the question is still what we understand by assessment. This is the real crux of the matter, about which there is a certain amount of disagreement among researchers. This is a stimulating issue which demands a solution for some of the following questions:

1. What do we understand by the learner's levels of skills and knowledge?
2. How do we categorise these concepts?
3. If ever there were agreement on the rating scales and everything were rationalised, even then the question would be the range of agreement / disagreement that could be considered acceptable among raters.

4.3. *The rating of the open question subtest and the essay*

The next step is to analyse the subjective components of the CUEE test shown in the following Table 5.

Table 5

Open question Subtest	Gender 1 male 2 female	Working place 1 University 2 Secondary school	Mean	Standard Deviation	Number
	1	1	6,5083	2,4856	150
		Total	6,5083	2,4856	150
	2	1	5,1917	3,1388	150
		2	4,1867	3,0373	150
		Total	4,6892	3,1241	300
	Total	1	5,8500	2,9023	300
		2	4,1867	3,0373	150
		Total	5,2956	3,0474	450
Essay	1	1	5,8878	2,2216	150
		Total	5,8878	2,2216	150
	2	1	5,6278	2,5525	150
		2	3,6945	2,6034	150
		Total	4,6611	2,7499	300
	Total	1	5,7578	2,3923	300
		2	3,6945	2,6034	150
		Total	5,0700	2,6474	450

Descriptive statistics. Open question subtest and essay

The difference between two means: open question items and essay are quite remarkable. In the open questions both male raters score higher than female raters, and University female raters higher than Secondary School female raters (UMR Mean = 6,5083, UFR Mean = 5,1917 and SSR Mean = 4,1867).

As for the essay the mean is almost the same for university male and female raters but it is very different with respect to Secondary School female raters (UMR Mean = 5,8878, UFR Mean = 5,6278 and SSR = 3,6945)

Two between-group factors: gender and working place are combined with the components of the subjective subtest: open question items and essay in the ANOVA for repeated measures. The F shows significant differences across the different components of the test: essay and open questions $F_{1,447} = 9.152$ and $p<0.01$. There are also interaction effects between these two components of the subtest and gender and working place, with $F_{1,447} = 14.361$ and $p < 0.001$ for gender and $F_{1,447} = 11,085$ and $p<0.001$ for the working place variable. Two principle effects between groups are described with $F_{1,447} = 8.051$ and $p<0.01$ for gender and $F_{1,447} = 27.962$ and $p<0.001$ for working place.

Obviously, it is expected that the significant differences of the subjective subtest continue in its components. Some light will be shed on the inter-rater scoring behaviour by studying the components in depth. The data show that female and male university raters have a similar rating behaviour in the essay and considerable differences in the open questions subtest. In the latter case the assessment of this behaviour could have

something to do with the pasting of chunks of words, that male raters might have failed to notice and female raters might have examined thoroughly.

The working place factor most clearly defines between the various rating groups in the essay correction. The performance of the University and Secondary School raters shows a big gap in their scoring procedure. This leads the inference that there may exist a divergence in their expectations about how the examinees are supposed to perform. This finding allows us only to claim, once more, that there are significant differences among these scoring groups in terms of the working place and that all tentative suggestions on the criteria that are used in different academic environments just mean new fields of research. Meanwhile, as we have not carried out an introspective study on what is going on in the raters' mind, we are not entitled to say whether knowledge of language took priority over content or which criteria had the greatest weighting in their assessment.

4.4. *The raters' scoring behaviour towards open questions vs essay and form vs content*

In phase II our aim is to see from another perspective - all the raters graded the same exam - if the significant differences in the large sample above studied are confirmed. In this stage, first of all, the components of the subjective subtest: open questions and essay together with form and content, two subcategories derived from the essay are considered, and then, the scoring behaviour of the raters towards a poor exam and a good exam.

Table 6

		Mean	N	Standard Deviation	Standard Error of the Mean
Pair 1	Essay	4,9483	100	2,2319	,2232
	Open Q.	6,4725	100	2,4749	,2475
Pair 2	Content	5,4567	100	2,4468	,2447
	Form	4,3350	100	2,3528	,2353

Descriptive statistics of correlated samples

As for the first pair there is correspondence between the data in Table 5 and the data presented above in Table 6. The statistics presented are similar to those in which raters marked different exams. Content average has higher weighting than form average in the essay. The correlation of the two subcategories introduced is higher ($r = 0.734$) than the corresponding to open question items and essay ($r = 0.566$).

The Friedman test presents significant differences across the different components with $\chi^2_3 = 94.503$ and $p<0.001$. As a "post hoc" test the Wilcoxon test is carried out in order to examine the paired differences: form / content. To maintain the level below 0.05

the Bonferroni's correction is applied. The test shows a significant difference of $z = -6.697$ and $p < 0.001$ between form and content. The rating of form was lower than that of content in 88 scores out of 100.

These statistical tests confirm the data of the large sample in which significant differences among raters are found, though they are not confirmed in terms of gender and working place because of the size of the sample. On the whole, broad versus narrow interpretations, strictness versus leniency on the grounds of possible different lexico-grammatical perspectives can be observed. Probably, it is a matter of seeing grammar as more prescriptive than descriptive among SSR on account of the curricula they work with throughout the year. Traditionally, grammatical studies had the goal of providing a relatively complete catalogue of the forms of language and description of the rules for combining these forms: inflectional morphology, derivational morphology, word order and various kinds of subordination. In prescriptive approaches, deviations from the norm are generally considered erroneous.

The different scoring behaviour on the content and form subcategories leads us to consider all raters closer to the prescriptive approach when dealing with form. This strictness leads us to think that they are more concerned with the elements on which communicative effectiveness is built than with the level of understanding that they generate. This tendency could be understood as a tendency among raters to penalise form, in which the grammatical elements are more specific, and to reward content, where criteria are less specific.

4.5. *The raters' performance depending on the quality of the exam*

As a final step to accept or reject that there are significant differences among raters in terms of gender and working place we have focused our attention on the raters' scoring behaviour in a exam considered poor and in another labelled a good one.

As for the range of scoring if we take content as a reference, the minimum score reaches 1.67 and the maximum 6.67 for subject 1, among the twenty raters, whereas for subject 5 the range goes from 3.33 to 10.

Table 7

	Working place	Number	Mean	Standard deviation	Standard Error of the mean
Open question	University	10	6,0000	1,2910	,4082
	Grammar School	10	4,5000	1,6874	,5336
Essay	University	10	3,6667	1,5316	,4843
	Grammar School	10	2,9333	1,0517	,3326
Content	University	10	4,7333	1,9566	,6187
	Grammar School	10	3,3500	1,6582	,5244
Form	University	10	3,0000	,8958	,2833
	Grammar School	10	2,1833	,8144	,2575

Working place. Subject n. 1 (poor exam)

Table 8

	Working place	Number	Mean	Standard deviation	Standard error of the mean
Open question	University	10	9,5000	,8740	,2764
	Grammar School	10	9,2500	1,0541	,3333
Essay	University	10	8,0833	1,1146	,3525
	Grammar School	10	8,4833	1,5040	,4756
Content	University	10	8,0000	1,5316	,4843
	Grammar School	10	8,3167	2,1088	,6669
Form	University	10	7,6667	1,1653	,3685
	Grammar School	10	8,0833	2,0050	,6340

Working place. N. 5 (good exam)

The sample is quite small and as the differences between the means are neither relevant to the poor nor to the good exam, we focus our study on the working place factor.

The working place has a different effect on the assessment of the variables we are dealing with. The Secondary School raters are stricter than the University ones when the exercise is poor, Subject 1, and the roles are slightly inverted when they are rating good exercises, Subject 5, as can be appreciated in Tables 7 and 8. Such different behaviour in assessing each exam leads to significant differences in the case of the poor student in the open question items ($t = 2.233$ and $p < 0.05$) and in the form ($t = 2.133$ and $p < 0.05$).

These results are in agreement with the comments previously made on the subcategories of the essay, showing this time strictness versus leniency between Secondary School and University raters respectively. The latter seem to be nearer descriptive rather than prescriptive grammatical interpretations. It is likely that the University academic environment leads them to be more concerned with communicative effectiveness than with grammar rules and to note the existence of grammatically correct alternatives (such as active versus passive voice) or to accept lexical items beyond the marking scheme established. But again these are tentative explanations of the raters' behaviour which deserve further research.

5. CONCLUSIONS

These findings lead us to claim that:

1. It would be advisable to pilot the Complutense University Entrance Exam to control the extraneous variables that can affect the rating of the objective subtest.
2. Secondary School raters are more concerned with scoring accuracy than University female raters, who are more lenient and probably more highly influenced by intelligibility than by errors.
3. As for gender rating behaviour among University raters, leniency is greater among men than among women both in the objective and in the subjective subtests.

4. Female and male university raters have a similar rating behaviour in the essay and significant differences in the open questions subtest. The scoring of this subtest opens new scope for research in terms of gender.
5. There are significant differences between objective and subjective subtests and between open questions and essay and in terms of the working place of the raters.
6. The second stage of this study confirms the data of the first stage in which significant differences across the different components of the test are found among raters.
7. Significant differences are noticeable in the rating of poor exams in terms of the working place while a levelling tendency can be observed if the exam is good.

These claims lead us to affirm that there are significant differences mainly in terms of working place in the CUEE and to propose new fields of research into the underlying criteria which guide the raters' assessment and narrow the wide range of scoring among raters.

5. ABBREVIATIONS

ANOVA: Analysis of variance.
CPE: Certificate of Proficiency in English.
CUUE: Complutense University Entrance Examination.
FCE: First Certificate in English.
ET: English test.
UMR: University male raters.
UFR: University female raters.
SSR: Secondary School raters.

NOTES

* This research has its origin in the project "Analysis on test's parameters" (ref: PR94-118), carried out at the Measurement and Computer Analysis Department in OISE, Toronto, Canada, with the financial support provided by the DGICYT. It is a development of an earlier paper presented by H. Herrera and M. Amengual at the Conference of AESLA XVIII, Barcelona, May 2000. My acknowledgements also to Mª Rosario Martinez Arias, Catherine Millar and Michael White for their helpful comments on the draft and to the anonymous referees for their patience and interest.

1. Length, legibility, grammar, structure, communicative effectiveness, tone, vocabulary, spelling, content, task realisation, punctuation. (Bachman 2000)
2. The use of ET henceforth will refer to the English Test in the Spanish University Entrance Examinations.

REFERENCES

Bachman, L.F. 1995. "Investigating Variability in Tasks and Rater Judgements in a Performance Test of Foreign Language Speaking". *Language Testing* 12, 2: 239-57.

Bachman, L.F. 2000. "Modern language testing at the turn of the century: assuring that what we count counts". *Language Testing* 17, 1: 1-42.

Baird, J.A. 1998. "What's in the Name?. Experiments with Blind Marking in A-Level Examinations". *Educational-Research* 40, 2: 191-202.

Berninger et al. 1996. "Assessment of planning, translating, and revising in Junior High Writers". *Journal of School Psychology* 34, 1: 23-52.

Brown et al. 1992. "Interactive learning environments: A new look at assessment and instruction", *Changing Assessments: alternative views of aptitude, achievement and instruction*. Eds. B. Gifford and M. O'Connor. Boston: Kluwer Academic.

Carson et al. 1990. "Reading-writing relationships in first and second language". *TESOL Quaterly* 24: 245-66.

Celce-Murcia, M. 1997. "Direct Approaches in L2 Instruction: A Turning Point in Communicative Language Teaching". *TESOL Quarterly* 31: 141-52.

Diederich et al. 1961. *Factors in Judgements of Writing Ability*. ETS Research Bulletin RB-61-15. Princeton, N.J.: Educational Testing Service.

Freedman, S. 1979. "How characteristics of student *essay*s influence teachers' evaluation" *Journal of Educational Psychology* 721: 328-338.

Gass, S. and E.M. Varonis. 1986. "Sex differences in NSS/NNSS interaction. *Talking to Learn: Conversation in Second Language Acquisition*. Ed. R.Day. Rowley, MA: Newbury House.

Gipps, C.V. 1994. *Beyond Testing: towards a theory of educational assessment*. London: The Palmer Press.

Gipps, C. and P. Murphy. 1994. *A Fair Test? Assessment, achievement and equality*. Milton Keynes: Open University Press.

Goddard-Spear, M. 1983. "Sex Bias in Science Teachers' Rating of Work". Contribution to the Second GASAT Conference, Oslo, Norway.

Grobe, C. 1981. "Syntactic maturity, mechanics and vocabulary as predictors of quality ratings". *Research in the Teaching of English* 15: 75-86.

Gyagenda, I.S. and G. Engelhard. 1998. "Patern, Domain, and Gender Influences on the Assessed Quality of Students Writing Using Weighted and Unweighted Scoring". Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998),

Hamp-Lyons, L. 1990. "Second language writing: assessment issues". Ed. B. Kroll. *Second Language Writing*. Cambridge: Cambridge University Press.

Hamp-Lyons, L. Ed. 1991. *Assessing Second Language Writing in Academic Contexts*. Norwood, N.J.: Ablex Publishing Corporation.

Hamp-Lyons, L.1995 "Rating nonnative writing: the trouble with holistic scoring". *TESOL Quaterly* 29: 759-62

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* 7: 89-107.

Herrera Soler, H. and M. Amengual Pizarro. 2000. "El lugar de trabajo y el género del corrector en la prueba de inglés de la Selectividad." Congreso de AESLA XVIII, Barcelona Mayo 2000.

Huot, B. 1990 "The literature of direct writing assessment: major concerns and prevailing trends". *Review of Education Research* 60: 237-63.

Jacobs et al. 1981. *Testing ESL Composition: a practical approach.* Rowley MA: Newbury House.

Pennington, M.C. and S. So. 1993. "Comparing writing process and product across two languages: A study of 6 Singaporean university student writers". *Journal of Second Language Writing* 2: 41-63.

Pica et al. 1989. "Comprehensible output as an outcome of linguistic demands on the learner". *Studies in Second Language Acquisition* 11/1: 63-90.

Raimes, A. 1990. "The TOEFL test of written English: causes for concern". *TESOL Quarterly* 24: 427-42.

Russikoff, K.A. 1994. "Hidden expectations: Faculty Perceptions of SLA and ESL Writing Competence". Paper presented at the Annual Meeting of TESOL, Baltimore, March 8-12, 1994.

Sasaki, M. 1999. "Development of an analytic rating scale for Japanese L1 writing". *Language Testing* 16,4: 457-478.

Shavelson, R.J. et al. 1992. "Performance assessments: political rhetoric and measurement reality". *Educational Researcher* 21, 4: 22-27.

Shehadeh, A. 1999. "Gender differences and equal opportunities in the ESL classroom". *ELT Journal* 53/4: 256-261.

Steward, M. and C. Grobe. 1979. "Syntactic maturity, mechanics of writing and teachers' quality ratings". *Research in the Teaching of English* 13: 207-15.

Vaugham, C. 1991. "Holistic assessment: What goes on in the rater's mind?". *Second Language Writing in Academic Contexts.* Ed. L. Hamp-Lyons. Norwood, N.J.: Ablex Publishing Corporation.