JOURNAL OF ENGLISH STUDIES - VOLUME 4 (2003-2004), 23-36

RATER DISCREPANCY IN THE SPANISH UNIVERSITY ENTRANCE EXAMINATION

MARIAN AMENGUAL PIZARRO
University of the Balearic Islands

ABSTRACT. *This study investigated the level of inter-rater and intra-rater reliability of thirty-two raters in the evaluation of the composition subtest of the English Test (ET) in the Spanish University Entrance Examination (SUEE). Raters were asked to evaluate ten compositions holistically on two different scheduled data collection sessions (PRE and POST). The results show that although there are no significant differences between the holistic PRE and POST scores, there exists a substantial discrepancy across raters in relation to consistency and harshness of scoring. On the contrary, results reveal that, in general, raters are self-consistent in the evaluation of compositions. On the basis of these results, and given the way SUEE scores will affect the career of a candidate, it is believed that much more emphasis should be placed on establishing an acceptable level of inter-rater reliability in order to ensure that SUEE results are as fair and consistent as possible.*

> It is well established that some raters are consistently harsher in their assessment of a candidate's ability than others […] From the point of view of the candidate, it becomes a matter of luck whether they are assessed by particular raters. A candidate may draw the most lenient member of the rating team and benefit as a result or, alternatively, s/he may draw the harshest member and may suffer the consequences of this. (Wigglesworth 1993: 305).

1. INTRODUCTION

In recent years, direct tests of writing have become standard practice for the assessment of second language (L2) writing abilities (Shohamy 1995; Tedick 1998). According to Cumming (1997), the writing of expository or narrative

compositions has now come to be considered an important indicator of students' English proficiency in second language academic contexts. From a communicative perspective to language instruction, compositions are considered to be realistic and communicative tasks. In fact, validity and good impact on teaching are some of the main advantages attributed to direct tests of writing such as compositions (Moss 1994).

At the same time, among different methods of evaluating compositions, holistic assessment has gained wide acceptance in second language testing. This method requires raters to make a single, global judgement based on the overall impression the text produces (Nagy *et al* 1988). However, its reliance on human interpretation has made holistic scoring a major area of controversy in writing assessment (Charney 1984; Cumming *et al* 2001).

The intrinsic weakness of direct tests of writing, therefore, lies in the fact that at all stages they are a wholly human endeavour (Hamp-Lyons 1990). This makes objective assessment of compositions very difficult. Such tests require subjective evaluations. Thus, most research emphasis has been placed on establishing the reliability of scoring among raters (i.e. inter-rater reliability) in order to show that compositions can be scored as fairly and consistently as possible (Weigle 1994; North and Schneider 1998; Lumley 2002).

However, fluctuations in scores associated with rater factors are extensive (Huot 1990; Lumley and McNamara 1995; Weigle 1998; Gamaroff 2000; Kondo-Brown 2002; Amengual 2003; 2004). Furthermore, raters are recognised to be one of the main sources of measurement error in assessing a candidate's performance (Milanovic *et al* 1996; Herrera 2001). To overcome the problem of rater reliability, many researchers advocate a further refining of the descriptive evaluative criteria and the procedures for scoring by means of training raters and maintaining consistent agreement among those raters (i.e. inter-rater reliability).

Rater reliability studies have not only addressed the issue of inter-rater reliability or agreement between different raters but also on intra-rater reliability or agreement between the same rater on the ratings assigned to compositions. Since trying to reconcile raters' subjectivity with objective precision seems extremely difficult to achieve, rater reliability has been defined as the greatest bugbear in assessment (Moss 1994; Gamaroff 2000).

Despite rater reliability concerns, the qualities and current status of written composition in evaluation practices has remained central to the English Test (ET) in the Spanish University Entrance Examinations (SUEE). Therefore, holistically oriented composition tasks are a conventional measure used to assess L2 writing in the ET. But are ET scores as reliable as they should be? This is one of the main

questions we should ask ourselves given the effect of those scores on students' lives and the social consequences derived from them. If we are to draw conclusions on the basis of the uses and interpretations of ET scores (Messick 1992) one of our major professional concerns should be to ensure that those scores are fair and, therefore, reasonably reliable (ILTA 2000; Alderson and Banerjee 2001).

1.1. The Test

The testing context for this study is the composition subtest of the English Test (ET) in the Spanish University Entrance Examination (SUEE). The ET is taken as a screening norm-referenced test to enter a Spanish university and according to Herrera (1999: 90) its main aim is to discriminate as reliably as possible. Therefore, the ET is expected to rank students according to their proficiency and spread students out into a normal distribution so that their performances may be compared.

The ET consists of four or five different subtests based on a reading passage. Although there might be some slight variations in the design of the ET across Spanish universities, the ET includes a composition subtest where students are required to perform a brief composition task (e.g. about 150 words) on the students' choice between two prompts related to the subject matter of the initial reading passage. The time allowed for the total completion of the ET cannot exceed an hour and a half.

Owing to practical obstacles such as the limited time that raters can devote to the marking of the ET, compositions are rated impressionistically by individual raters. A common set of evaluative criteria is used by raters but they are so flexible and open that "each rater comes to rely on his own method" (Vaughan 1991: 121).

Raters are expected to score examinations within the first five days after the date of the ET. They are qualified teachers working either at the University or in Secondary Education. However, they have not received any training for large scale-assessment.

Even so very little is known about the level of inter-rater reliability in the evaluation of compositions in the ET. This is a very strange and worrisome state of affairs given the way SUEE results will affect students who want to gain entrance to the faculty of their choice.

2. PURPOSE

The central purpose of this study is to investigate the degree to which differences among raters (i.e. inter-rater reliability) exist in the composition scores

of the ET. It also explores differences in rater consistency within the same rater (i.e. intra-rater reliability). To investigate this latter point we will use a pre- and post- design. Specifically, the following research questions were posed:

1. What level of agreement exists among all raters in relation to the total composition scores (i.e. inter-rater reliability)?
2. Is there a significant difference between the average total composition scores assigned by raters?
3. What level of agreement exists within the same rater on different occasions (PRE and POST) (i.e. intra-rater reliability)?

3. METHOD

3.1. SUBJECTS

Thirty-two raters participated in this study. In selecting the raters, consideration was given to their professional background and experience in marking the SUEE compositions. All of them had participated in previous SUEE administrations and were qualified University and Secondary Education teachers. Following the SUEE policy, none of the raters had followed any training program for evaluating ESL compositions.

3.2. MATERIALS

Ten compositions were randomly selected from a pool of 136 compositions from a previous administration of the SUEE, all of them written on the same prompt for the sake of comparability. The reading passage which formed the basis of the composition was a brief description of the city of London. The actual essay topic was *Write a composition of 100-150 words on the following topic: a holiday in London mentioning the places you would visit and why.* Each of the original compositions was numbered from 1 to 10. They were then photocopied and given to each rater in exactly the same order, so as to avoid any effect for order of presentation on marking behaviour.

3.3. PROCEDURES

All raters took part in two individually scheduled data collection sessions, as follows:

1) One in February 2000 (PRE).

2) One in May 2000 (POST).

These data collection sessions are described briefly below.

a. *PRE*: Raters were asked to score the same 10 SUEE essays holistically on a scale from 1 to 10 (whole numbers only) in the same prescribed order of presentation. No specific marking criteria were given. The raters were asked to do the task as if they were evaluating the ET compositions for the SUEE. Raters took from 1 to 2 weeks to correct the essays.

b. *POST*: Three months later, all raters were given the same set of ten compositions they had rated in the PRE data collection session although they were not informed of this fact. In fact, some raters did not remember having read the essays before. Raters were again required to assign a holistic score to the compositions on the same scale (from 1 to 10, whole numbers only). It is important to note that this time compositions had been arranged in a different random order of presentation, so as to avoid contamination of previous results. Raters were then asked to rate the compositions in the same prescribed order. The order of presentation for the essays in the PRE and POST data collection sessions was the following:

PRE:	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
POST:	C10	C8	C3	C7	C9	C5	C2	C1	C6	C4

Cn refers to composition 1, composition 2…etc.

Composition results were collected within one to two weeks after the beginning of the POST collection session. The specific purposes and details of this research were not revealed until all the ratings had been completed.

4. RESULTS AND DISCUSSION

The results for the two research questions listed earlier are reported below:

4.1. Are holistic scores reasonably reliable?

In order to answer the first research question, a simple comparison was firstly made of the holistic scores awarded by the 32 raters to compositions in the PRE and POST sessions respectively. Table 1 below summarises the descriptive statistics obtained:

	N	Range	Minimum	Maximum	Mean	Standard deviation
Holistic-PRE	32	4.10	2.90	7.00	4.85	1.08
Holistic-POST	32	2.80	2.80	5.60	4.55	0.67

Table 1. *Descriptive statistics: Holistic scores PRE and POST.*

As it can be seen from the data, the mean global scores is found to be slightly higher in the PRE ($\overline{x}$ = 4.85) than in the POST ($\overline{x}$ = 4.55) session. However, there is far more variability among raters in the PRE scores (SD = 1.08) than in the POST scores (SD = 0.67), as indicated by the higher standard deviation figure in the PRE session. In other words, scores tend to be less homogeneous and therefore less consistent in the PRE than in the POST session. The range of scores is also wider in the PRE (4.10) than in the POST (2.80).

It should be stressed here that in the context of the SUEE a difference of 3 and 4 points across compositions on a 10-point scale might be considered rather extreme, especially if we consider the dramatic consequences those results might have on students' lives, as we have pointed out above.

The first correlation statistic employed in this study was the standard parametric Pearson correlation which was calculated for the average composition totals (PRE and POST) assigned by raters (Table 2). Although correlations were found to be significant at the 0.01 probability level, the Pearson correlation figure (r = 0.508), indicates that, in general, the amount of agreement among raters is rather low.

Pearson		Holistic PRE	Holistic POST
Holistic PRE	Correlation coefficient Sig. (unilateral)	1.000	0.508**
Holistic POST	Correlation coefficient Sig. (unilateral)	0.508**	1.000

** $p < 0.01$ (unilateral).

Table 2. *Holistic scores correlations (PRE and POST).*

Reliability was further addressed by exploring the intra-class correlation to determine the degree of relationship of overall scores among raters. Table 3 below

shows the intra-class correlation estimate for the essay totals of all raters in both occasions (PRE and POST). The figure obtained here (ρ_I = 0.6556), which is a truer measure of actual agreement than the Pearson correlation figure, is slightly more encouraging. However, following Fleiss[1] criteria (1986), an intra-class correlation of 0.6556 is still rather modest overall. On the basis of these correlation figures (both *r* and ρ_I), it seems that the overall level of agreement among raters is not particularly strong. Furthermore, according to Hatch and Lazaraton (1991: 534) a correlation lower than 0.75 in cases like this one where ratings were given without any predetermined evaluative criteria or training would indicate that raters had applied dramatically different evaluative criteria.

Consistency level
Intra-class correlation mean = 0.6556
95.00% C.I.: Lower = 0.2944 Upper = 0.8319
F= 2.9033 DF = (31, 31, 0) Sig. = 0.0020 (Test value = 0.0000)
Reliability Coefficient
N of cases = 32
α = 0.6556

Table 3. *Intra-class correlations for the composition totals (PRE and POST) of all raters.*

4.2. Is there a significant difference between the average composition scores assigned by raters on both occasions (PRE and POST)?

A paired *t* test on the differences between the average composition PRE and POST totals was carried out to examine this question. The results, as shown in Table 4 below, indicate that there are not significant differences between the holistic PRE and POST scores (t = 1.84, p = 0.075). Thus, despite the wide spread of scores around the mean (SD = 9.1) the differences between the average global ratings are not significant. However, it is worth noting that this result is on the threshold of significance at the 0.05 probability level. This suggests that this result may be attributable, in part, to the small size of the data set.

1. According to Fleiss criteria (1986) a correlation <0.40 indicates a low level of agreement; a correlation of 0.41-0.75 indicates a regular-good level of agreement and a correlation of 0.76-1.00 indicates a clearly high level of agreement.

	Differences							
	Mean	Std. Deviation	Std. Error Mean	Confidence 95% Interval of the Difference Lower	Upper	t	df	Sig.
Pair HPRE-HPOST	2.9688	9.10684	1.60988	– 0.314	6.2521	1.844	31	0.075

*HPRE and HPOST refer to Holistic PRE scores and holistic POST scores respectively.

Table 4. *T-Tests for paired samples.*

4.3. What level of agreement exists between the same raters on PRE and POST scores?

Table 5 shows the frequencies of point differences between PRE and POST scores for each rater. As it can be seen from the table, the majority of raters' POST scores (n = 24) were within 0 and 1 points of their corresponding PRE scores. Furthermore, the average difference between the PRE and POST scores of all raters was less than one point (0.7). These data show that, in general, the level of agreement between the same raters on both occasions (PRE and POST) is quite high. That is to say, raters seem to be self-consistent, the most crucial quality in a rater according to Wiseman (1949) and Oller (1979).

Points / Raters	0	1	2	3	4	Total Mean difference (PRE-POST)
R1		1				1.8
R2		1				0.5
R3		1				1.9
R4	1					0.7
R5	1					-0.3*
R6	1					0.9
R7	1					-2
R8				1		-0.4

R9		1				0.6
R10		1				0.4
R11				1		-0.5
R12		1				0.7
R13		1				0.1
R14	1					0.6
R15			1			-0.8
R16	1					-0.2
R17			1			-0.3
R18		1				1.7
R19	1					0.8
R20		1				1.6
R21		1				1.3
R22		1				-1
R23					1	-0.4
R24	1					0.1
R25		1				-0.3
R26		1				-0.4
R27			1			1.9
R28		1				-0.2
R29				1		-0.2
R30	1					0.3
R31	1					0.7
R32			1			-0.1
TOTAL	10	14	4	3	1	23.7

* The negative sign is due to the fact that scores were higher in the PRE than in the POST session.

Table 5. *Point differences between PRE and POST ratings of compositions.*

However, four raters (R15, R17, R27 and R32) were identified as having a PRE-POST rating difference of 2 points on a 10-point scale and four other different raters registered a PRE-POST rating difference of three and four points (R8, R11 R29 and R23) on the same scale. R23, in particular, was the less consistent since her scores were four points away from the PRE scores on the same compositions. This latter fact is a cause for concern in the context of the SUEE where the ET scores will

provide a qualification that will affect the career of a candidate, as it has already been mentioned above.

A scatterplot was used here to help identify the behaviour of each rater in relation to the mean. From the shape and slope of the ellipse, we can see that there is some evidence of a positive relationship between the PRE and POST scores. That is, as the candidates' scores on PRE increase so do their scores on POST.

However, the diagram also reveals certain dispersion of scores around the mean. This indicates some discrepancies in the behaviour of raters between the PRE and POST scores. R27 and R13 are identified as *outliers*. R27 seems to be extremely lenient since his average score is above 7 points on a 10-point scale. R13, on the contrary, seems to be extremely harsh since her average score is lower than 3 points on the same scale. An estimate of the reliability and consistency of the rater's scores would, perhaps, be useful in determining whether the rater should participate satisfactorily on future SUEE rating process occasions.

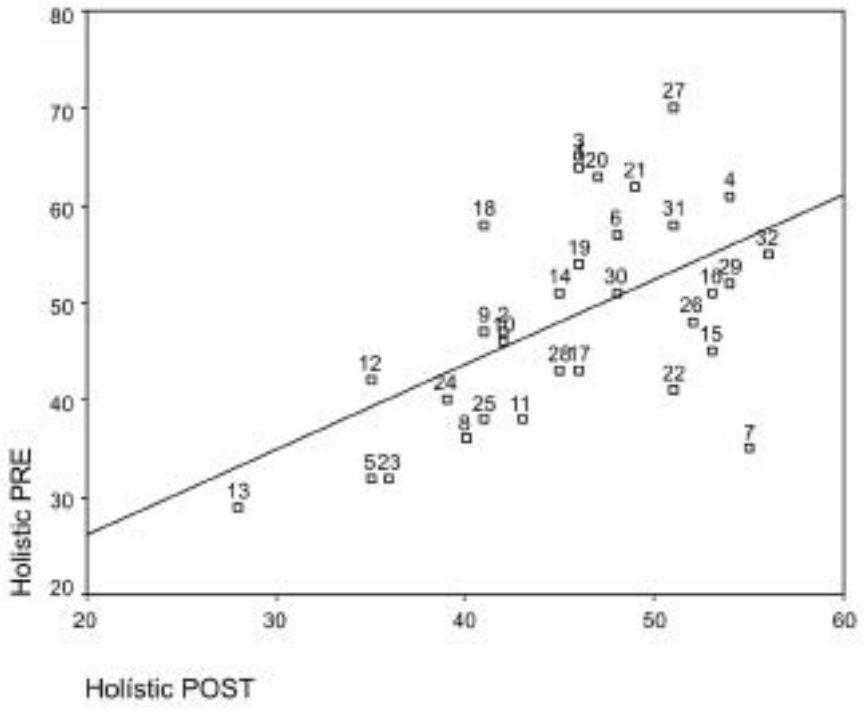

Figure 1. *Scatterplot of PRE and POST holistic scores.*

5. CONCLUSION

While both the Pearson correlations and the intra-class correlations for the composition totals of all raters are all significant at the 0.01 probability level (see Tables 2 and 3), they are still rather modest overall. This suggests that the level of agreement among raters is not particularly strong. Thus, it seems that raters have different rating styles and can greatly differ in their assessment of compositions.

The *t* test results (see Table 4) also reveal that there are no significant differences between the holistic PRE and POST scores. However, any conclusions drawn from these results must be regarded with caution since individual differences among raters are ironed out in the averaging process. In fact, the wide range of scores among compositions (see Table 1) indicate that there are important differences in the behaviour of raters in relation to consistency and harshness of scoring. The variability of scores across raters that has been discovered in this study should lead us to call into question the practice adopted by the SUEE of basing judgements of candidates on single ratings, particularly if we consider the dramatic personal consequences for students, who may be unfairly deprived of gaining entry to the faculty of his/her choice, which depends on these results.

With regard to intra-rater reliability, it is interesting to note that the level of agreement of each rater on both occasions (PRE and POST) is quite high. Despite some exceptions, it seems that raters, in general, are internally consistent. But then, there are still some outliers or extreme cases where "we cannot be sure that the rater will not mark differently before breakfast (a good or bad one) than after." (Gamaroff 2000: 45).

Although it is evident from the data that the rating process is complex, it also seems clear that it is difficult to derive consistent results from untrained raters due to the large randomness associated with their scores (Weigle 1994). If reliability of scores concerns us and we discover that our rating behaviour is fairly inconsistent we should take steps to redress this situation. Training raters may be a good solution to modify and improve evaluation practices. Failing that, the traditional technique of double ratings with an averaging of raw scores seems amply justified (Wigglesworth 1993; Weigle 1994; Lumley and McNamara 1995). However, there is no room for this assessment technique in the context of the SUEE due to the increase in economic costs involved in the process. A more feasible alternative which requires little additional time and expense would include, for example, the refinement of the evaluation criteria so as to modify rater expectations and raise awareness of the need for inter-rater agreement.

The results and conclusions of this study are naturally tentative. However, it is hoped that this research will allow us to reconsider the central importance of rater consistency in evaluation marking and can serve as a model for further research on the design of techniques addressed to eliminate undue influences on scores.

REFERENCES

Alderson, J. C., and J. Banerjee. 2001. "Language Testing and assessment (Part 1)". *Language Teaching* 34: 213-36.

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

—, 2004. "Análisis de la fiabilidad en las puntuaciones holísticas de ítems abiertos". Published PhD thesis. CERSA: Complutense de Madrid University.

Charney, D. 1984. "The validity of using holistic scoring to evaluate writing: a critical overview". *Research in the Teaching of English* 18: 65-81.

Cumming, A. 1997. "The Testing of Writing in a Second Language". *Encyclopedia of Language and Education: Language Testing and Assessment* (7). Eds. C. Clapham and D. Carson. 51-63.

Cumming, A., R. Kantor, and D. Powers. 2001. *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making, and development of a preliminary analytic frame-work.* TOEFL Monograph Series. Princeton, NJ: Educational Testing Service.

Fleiss, J. L. 1986. *The Design and Analysis of Clinical Experiments.* New York: WILEY.

Gamaroff, R. 2000. "Rater Reliability in Language Assessment: the Bug of all Bears". *System* 28: 31-53.

Hamp-Lyons, L. 1990. "Second Language Writing: Assessments Issues". *Second Language Writing.* Ed. B. Kroll. Cambridge: Cambridge University Press. 69-87.

Hatch, E., and A. Lazaraton. 1991. *The Research Manual: Design and Statistics for Applied Linguistics.* New York: Newbury House.

Herrera, 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.

—, 2001. "The Effect of Gender and Working Place of Raters on University Examination Scores". *Revista Española de Lingüística Aplicada* 14: 161-79.

Huot, B. 1990. "The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends". *Review of Educational Research* 60: 237-63.

ILTA (International Language Testing Association) 2000. *Code of Ethics for ILTA*. Available on-line at http://www.Dundee.ac.uk/languagestudies/Itest/ilta/ilta.html (accessed February 2003).

Kondo-Brown, K. 2002. "A FACETS analysis of rater bias in measuring Japanese second language writing performance". *Language Testing* 19: 3-31.

Lumley, T. 2002. "Assessment criteria in a large-scale writing test: what do they really mean to raters?" *Language Testing* 19: 247-76.

Lumley, T., and T. F. McNamara. 1995. "Rater characteristics and rater bias: implications for training". *Language Testing* 12: 54-71.

Milanovic, M., N. Saville, and S. Shen. 1996. "A Study of the Decision-making Behaviour of Composition Markers". In *Performance, Testing, Cognition and Assessment*: Selected Papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem. Eds. M. Milanovic and N. Saville. Cambridge: University of Cambridge Local Examinations Syndicate and Cam-bridge University Press.

Messick, S. 1992. "Validity of Test Interpretation and Use". In *Encyclopedia of Educational Research*. Sixth edition. Ed. M. C. Alkin. New York: Macmillan. 1487-95.

Moss, P. 1994. "Can There Be Validity without Reliability?" *Educational Researcher* 23: 5-12.

Nagy, P., P. Evans, and F. Robinson. 1988. "Exploratory Analysis of Disagreement among Holistic Essay Scores". *The Alberta Journal of Educational Research* 4 (XXXIV): 355-74.

North, B., and G. Schneider. 1998. "Scaling descriptors for language proficiency scales". *Language Testing* 15: 217-62.

Oller Jr, J. W. 1979. *Language Test at School*. London: Longman.

Shohamy, E. 1995. "Performance assessment in language testing". *Annual Review of Applied Linguistics* 15: 188-211.

Tedick, D. J., ed. 1998. *Proficiency-oriented language instruction and assessment: a curriculum handbook for teachers*. Minneapolis, MN: Center for Advanced Research on Language Acquisition, University of Minnesota.

Vaughan, C. 1991. "Holistic Assessment: What Goes on in the Rater's Mind?". *Assessing Language Writing in Academic Contexts*. Ed. L. Hamp-Lyons. Norwood, NJ: Ablex. 111-25.

Weigle, S. C. 1994. "Effects of training on raters of ESL compositions". *Language Testing* 11: 197-223.

—, 1998. "Using FACETS to model rater training effects". *Language Testing* 15: 263-87.

Wigglesworth, G. 1993. "Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction". *Language Testing* 10: 305-36.

Wiseman, S. 1949. "The marking of English composition in English grammar school selection". *British Journal of Education Psychology* 19: 200-09.