EUROPEAN ACADEMIC RESEARCH
Vol. III, Issue 4/ July 2015

Impact Factor: 3.4546 (UIF)
DRJI Value: 5.9 (B+)

ISSN 2286-4822
www.euacademic.org

The Question of Reliability in Script Scoring Practices

SAMAR KAMAL FAZLI
Assistant Professor of English
Department of Humanities, CIIT Lahore
Pakistan

Abstract:

The present study analyzes the evaluation system for the compulsory subject of English at undergrad level in Bahauddin Zakariya University, Multan. The study considered how different evaluators assign scores to the same writing tasks in certain answer scripts. Mainly, the study concentrated on the reliability of scoring with reference to the paper A of the annual examination conducted for the candidates of Arts subjects at graduation level. Paper 'A' was chosen because it contained descriptive and subjective type questions and these types of questions are prone to subjective marking. This study has been conducted with the help of randomly selected university approved examiners and answer scripts. A total of 50 scripts were marked by five examiners one after the other. This data was analyzed using the SPSS software in order to apply the ANOVA technique analysis. The analyses were made question wise as well as on the total marks awarded by the five examiners to the fifty scripts. The ANOVA technique was applied on the data in order to compare the performance of all the five examiners. The results indicated that there was no significant difference in the marking of all the five examiners on the total marks awarded. However, small differences were recorded in the marking of the individual questions.

Key words: evaluation system, script scoring practices, reliability, ANOVA

1. Introduction

English is one of the major subjects taught at all levels of education from grade one to the graduation level all over Pakistan. English is also the subject that has a high rate of failures at all levels especially at the graduation level. Students, who fail to achieve a passing grade in English, often accuse the evaluation system for the subject of English. It is a common complaint by the candidates appearing in the compulsory paper of English at all levels that the scripts of English are not marked properly by the raters. This complaint is more noticeable among the graduate level students. There is another view that the percentage of failure and/or low scores at graduation level is higher in the compulsory subject of English than in other subjects. The present study was carried out to address this complaint empirically. In this study, an attempt is made to measure the average score awarded to fifty scripts by five different examiners. If the difference is higher, the complaint of the candidates is valid and must be taken account of. On the other hand, if the difference is not high enough, the complaint is not much valid and the scoring of the scripts is objective and reliable to a great extent.

2. Literature Review

Scoring of the scripts is highly critical in the process of evaluation because marks are ultimately used in awarding grades and divisions to the candidates who take exams. A score in a writing assignment is the outcome of an interaction that involves not merely the test taker and the test, but the test taker, the prompt or the task, the written text itself, the rater(s) and the rating scale according to Hamp Lyons [1] and McNamara [2]

When several teachers mark the same student's work they score it differently. In academic development programmes, too, there are often widely differing values about standards and marking criteria among those academics present Brown, Rust and Gibbs [3]. According to Karmal [4], passing or failing in an examination could vary according to who marked a particular paper.

In an investigation, Falls [5] asked 100 English teachers to grade a paper already evaluated by a committee as excellent. The 100 teachers did not know that a teacher had previously evaluated the paper. He reported that the grades on this paper were between 60 and 98.

In contexts where essays are marked by more than one marker, discrepancies often exist among the different raters as in Hamp Lyons [6], Lee [7], Vann, Lorenz& Meyer [8], and Weir [9]. According to researchers, the scorer reliability is the key concept in test reliability Hughes [10]; Lumley [11]. Cho [12] believes that "rating discrepancy between raters may cause a very serious impediment to assuring test validation, thereby, incurring the mistrust of the language assessment process itself". Hughes Arthur [13] says about the scorer reliability that it is possible to quantify the level of agreement given by different scorers on different occasions by means of a scorer reliability coefficient. He further says, "While the perfect reliability of objective tests is not obtainable in subjective tests; there are ways of making it sufficiently high for test results to be valuable."

Researchers have also considered factors associated with scoring practices and hence with the scorers reliability. Alderson & Bachman [14], while acknowledging the difficulties raters face in assessing the essays, consider writing to be one of the most difficult areas of language to assess. Scorers may come from different backgrounds, may have different systematic tendencies like rater severity/leniency as said by Wiseman [14], may have different attitudes to errors as said by Connors & Lunsford [15] and by Lunsford & Lunsford [16], and may have different expectations of good writing as in Coffman [18] and Cho [12]. Payne [19] suggested that one method of increasing reliability of an essay test is to increase the number of questions and restrict the extensiveness of the answers. Also, language testing professionals e.g. Alderson [20], Hughes [10], Weir, [21] suggest double marking in order to achieve inter-rater reliability. Research also supports the view that scorer's reliability can be improved considerably by training the markers (Charney, [22]; Cho [12]; Douglas, [23], Huot, [24]; Weigle, [25] though it cannot completely eliminate the element of subjectivity Kondo-Brown [26]; Weir [21]; Wiseman [15]. For assessing the reliability of scores, Weigle [27] suggests that one of the important ways to investigate the reliability of the scores is to investigate the inter-rater reliability. In her words:

> " ….inter-rater reliability refers to the tendency of different raters to give the same scores to the same scripts."

3. Research Methodology

For assessing the reliability of scores, the researcher followed the method designed by Weigle [27]. According to her, one of the most important ways to investigate the reliability of the scores is to investigate the inter-rater reliability. In her words, inter-rater reliability refers to "the tendency of different raters to give the same scores to the same scripts." Further, she explains, "A complementary approach to investigate inter-rater reliability, particularly when more than two raters are involved, is through the analysis of variance (ANOVA). ANOVA can be used to compare the distribution of scores given by a set of raters (assuming they have all scored the same scripts). The two main statistics used to describe the distribution of scores are the mean, or average score, and the standard deviation, or the average amount that scores differ from the mean. ANOVA can be used to determine whether there is any statistical difference between the mean scores of raters, irrespective of the correlation among raters' scores." Keeping this design in view, the researcher decided to use ANOVA technique for investigating the reliability of scoring among different scorers/raters.

The researcher selected Paper "A" of Compulsory English administered at Graduation level by Bahaud Din Zakaria University, Multan for the study. This paper contains essay type/descriptive questions that are usually prone to subjective marking.

Fifty answer scripts were selected randomly and five evaluators were selected randomly out of a list of 60 on the panel of evaluators who mark the scripts of English for the Bahauddin Zakariya University, Multan. Each script was marked by all the five raters independently without knowing what scores have been assigned to the writing by other raters. Afterwards, ANOVA technique was used for investigating the reliability of scoring among different scorers/raters.

4. Data Analysis

Scoring, done by 5 different evaluators, is a quantitative variable used in the analysis. The examiners who evaluated the answer scripts were independent i.e. the scoring of the 1st evaluator did not affect that of the 2nd evaluator and so on.

Analysis of Question 1 shows that the average marks of question 1 given by the first examiner are 8.22 with a standard deviation of 4.42. Average marks given by the second examiner are 7.98 with a standard deviation of 5.09; by the third examiner are 10.76 with standard deviation of 6.10; by the fourth examiner are 7.16 with a standard deviation of 4.96; and by the fifth examiner are 5.76 with a standard deviation of 4.38. Multiple Comparison Test (Post Hoc Test) was used to investigate that average marking of which 2 examiners is significantly different. It showed that the examiner 1 and 3, 1 and 5, 2 and 3, 2 and 5, 3 and 4 and 3 and 5 differ significantly as their observed p Values are 0.012, 0.015, 0.006, 0.028, 0.000 and 0.000 respectively.

Analysis of Question 2 depicts that the average marks of question 2 given by the first examiner are 7.30 with a standard deviation of 2.70; by examiner 2 are 7.40 with a standard deviation of 2.91; by the examiner 3 are 6.95 with a standard deviation of 3.17, by examiner 4 are 8.90 with a standard deviation of 3.56, and by examiner 5 are 6.52 with a standard deviation of 3.44. The output of Post Hoc Test, shows that the examiner 1 and 4, 2 and 4, 3 and 4, 4 and 5, differ significantly as their observed p Values are 0.025, 0.035, 0.006 and 0.001 respectively.

Analysis of Question 3 shows that the average marks of question 3 given by the first examiner are 4.41 with a standard deviation of 1.98; by examiner 2 are 4.54 with a standard deviation of 2.15; by the examiner 3 are 4.5 with a standard deviation of 1.96; by examiner 4 are 5.16 with a standard deviation of 2.35; and by examiner 5 are 4.12 with a standard deviation of 2.37.

Since the calculated P value 0.2 is greater than 0.05, it means that the Null Hypothesis i.e. the average marking of 5 examiners is equal on the basis of Q3 is valid.

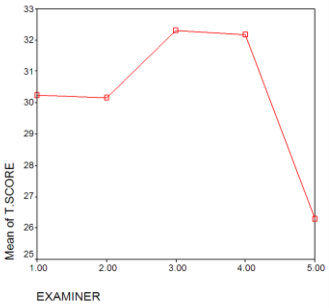
The average marks of Question 4 given by the first examiner are 4.77 with a standard deviation of 1.86; by examiner 2 are 4.93 with a standard deviation of 1.89; by the examiner 3 are 4.90 with a standard deviation of 1.89; by examiner 4 are 5.62 with a standard deviation of 2.22; and by examiner 5 are 4.70 with a standard deviation of 2.22. Since the calculated P value 0.218 is greater than 0.05, it means that we accept Null Hypothesis i.e. there is no significant difference in the average marking of the five examiners on the basis of Q4.

Analysis of Question 5 depicts that the average marks awarded by the first examiner are 4.34 with a standard deviation of 2.20; by examiner 2 are 4.37 with a standard deviation of 2.38; by the examiner 3 are 4.23 with a standard deviation of 2.02; by examiner 4 are 4.54 with a standard deviation of 2.58; and by examiner 5 are 4.06 with standard deviation of 2.48. Since the calculated P value is 0.903 greater than 0.05, It means that again the Null Hypothesis i.e. there is no significant difference in the average marking of the 5 examiners on the basis of Q5 holds true.

Analysis of Question 6 portrays that the average marks of question 6 given by the first examiner are 4.49 with a standard deviation of 1.80, by examiner 2 are 4.56 with a standard deviation of 2.12, by the examiner 3 are 4.43 with a standard deviation of 1.97, by examiner 4 are 4.89 with a standard deviation of 2.83, and by examiner 5 are 4.46 with a standard deviation of 2.41. Since the P value of Levene's is 0.002, which is less than 0.05, it means that, the Null Hypothesis stands rejected and population variances are not equal. In such a case, ANOVA cannot be applied. As a solution, a Non-parametric test i.e., Kruskal Wallis Test (H test), which is a parallel technique of ANOVA, has been applied. The P Value of this test is 0.657, which is greater than 0.05. It means that the average marking of Q6 by the five examiners does not differ significantly.

ANOVA Applied on the Total Scores Awarded by Each Examiner

Means Plots



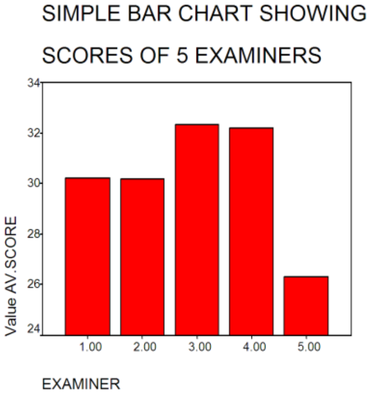
In this regard, the average scoring and standard deviation of 5 evaluators was calculated. For this purpose, the data was put into SPSS Software and the following results were obtained (shown in the following table).

Table

Examiners	**N**	**Mean**	**Standard Deviation**
1	50	30.24	15.25
2	50	30.16	16.36
3	50	32.32	16.22
4	50	32.19	17.72
5	50	26.30	16.45
Total	250	30.24	16.43

It is obvious that the average scoring of the first evaluator is 30.24, the second evaluator is 30.16, the third evaluator is 32.32, the fourth evaluator is 32.19, and the fifth evaluator is 26.30 with standard deviations of 15.25, 16.36, 16.62, 17.72, and 16.45respectively.

The above analysis can also be depicted by using a Simple Bar Chart.



It can be easily observed that average scoring is approximately close to each other. Although average scoring of the fifth evaluator is 26.30, which is different from the other 4 evaluators, it is not significant. This is because standard deviation of the 5th evaluator is 16.45, which is closer to the standard deviation of other evaluators. From this we can deduce that the average marking of the five examiners is to a large extent objective.

For this, the hypotheses were checked

Ho= the population variance of all the five evaluators is equal.

Hi: The population variance of all the five evaluators is not equal.

Level of significance, $\alpha = 5\%$

P-Value of Levene's Test = 0.677

Since the calculated P-value of Levene's Test is 0.677, which is greater than .05, means that we are going to accept Ho, i.e. the population variance of the five evaluators is approximately the same.

Now, a gateway to use ANOVA test is available as our key assumptions are fulfilled. Now we can use ANOVA.

For ANOVA, first of all, we state our Null and Alternative Hypotheses

H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

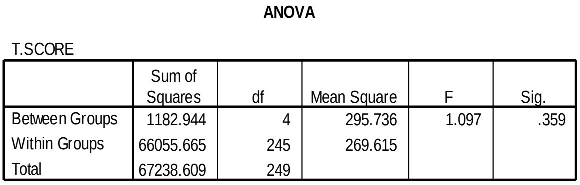
H_1: At least one mean is different.

Taking 5% risk, we have

$\alpha = .05$

Now again, the data was put into SPSS and the following results were obtained:

Table



ANOVA

T.SCORE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1182.944	4	295.736	1.097	.359
Within Groups	66055.665	245	269.615		
Total	67238.609	249			

P-Value of ANOVA Test= 0.359

As the P-value of ANOVA is 0.359, which is greater than 0.05, means, that Ho is to be accepted i.e.

"The Average Scoring of the Five Evaluators is almost same."

5. Findings

The results of our data analysis clearly show that there is no significant difference in average scoring of different evaluators. The major findings of the present study are given below:

For the question no.1, the examiner 1 and 3, 1 and 5, 2 and 3, 2 and 5, 3 and 4 and 3 and 5 differ significantly as their observed p Values are 0.012, 0.015, 0.006, 0.028, 0.000 and 0.000 respectively.

Similarly, for the question no.2, the examiner 1 and 4, 2 and 4, 3 and 4, 4 and 5, differ significantly as their observed p Values are 0.025, 0.035, 0.006 and 0.001 respectively. Whereas, no significant difference was found in the average marking of the five examiners in respect of Q3, Q4, Q5, and Q6.

The performance of all the five examiners based on the total scores awarded to the fifty answer scripts shows that the average scoring of all the five examiners is approximately similar. Although, average scoring of the Fifth evaluator is 26.30, which is a bit different from other 4 evaluators, yet it does not matter much. This is because standard deviation of 5th evaluator is 16.45, which is closer to the standard deviation of other evaluators.

6. Critical Insights

On the basis of our findings, we can make the following conclusions:

The charge against the evaluation process that it does not give reliable results is not true as different evaluators awarded almost the same scores to the same answer scripts. Therefore, the scoring of the answer scripts is reliable in this case. The researcher also traced the reasons of the sameness of the results. It was found out that although most of the evaluators mark the answer sheets using the general impression technique, as the evaluators are not given any scoring rubric, most of them devise their own rating scale to mark the papers. They even do not adhere to such a rating scale fully.

The following suggestions are being given to make the scoring of the paper of English more reliable:

Before the formal marking session starts, a search should be made to identify the 'benchmark' scripts which typify key levels of ability on each writing task. Copies of these then be presented to the scorers for initial scoring. Only when there is an agreement on these benchmark scripts, should scoring begin. Each task of each candidate should be scored independently by two or more examiners (as many examiners as possible should be involved in the assessment of each student's work), the scores being recorded on separate sheets. A third, senior member of the team should collate scores and identify discrepancies in scores awarded to the same piece of writing. Where these are small, the two scores can be averaged; where they are larger, senior member of the team will decide the score. It is also worth looking for large discrepancies between a candidate's performances on different tasks. These may accurately reflect his or her performance, but they may also be the result of inaccurate scoring.

Multiple scoring ensures scorer reliability, even if not all examiners are using quite the same standard. Nevertheless, once scoring is completed, it is useful to carry out simple statistical analyses to discover if anyone's scoring is unacceptably aberrant. Random reviewing of the scripts, after marked once, should be encouraged and practiced.

Before marking the scripts, the question paper should be discussed by the sub-examiners and the head-examiners. The examiners should be properly trained. Only those examiners should be appointed who have at least 5years of teaching experience at the graduation level. Moreover, the instructions about marking should be written and circulated to all the sub-examiners.

REFERENCES

[1] Hamp-Lyons, L. Second language writing: Assessment issues. Second language writing: Research insights for the classroom, 69-87. (1990).

[2] McNamara, T. F., & Candlin, C. N. Measuring second language performance (p. 165). London: Longman. (1996).

[3] Brown, S., Rust, C., & Gibbs, G. Strategies for diversifying assessment in higher education. Oxford: Oxford Centre for Staff Development. (1994).

[4] Karmal, L.T. and M.O. Karmal. Measurement and Evaluations in the Schools. McMillan Publishing Co. Inc., New York. 1978.

[5] Falls, James D. Research in Secondary Education. Kentucky School Journal. Pp 6, 12-46, Mehrens & Lehmann. (1928).

[6] Hamp-Lyons. L. Raters respond to rhetoric in writing. In H.W. Dechert and M. Raupach (Eds.). Interlingual Process (pp. 229-244). Tübingen: Narr. (1989).

[7] Lee, Y. An investigation into Korean markers' reliability for English writing assessment. English Teaching, 53(1),179-200.(1998).

[8] Vann, R.J., Lorenz, F.O., & Mayer, D.M. Error gravity: Faculty response to errors in the written discourse of nonnative speakers of English. In L. Hamp-Lyons (Ed.). Assessing second language writing in academic contexts (pp. 181-195). Norwood, NJ: Ablex Publishing Corporation. (1991).

[9] Weir, C.J. Understanding and developing language tests. New York: Prentice Hall. (1993).

[10] Hughes, A. Testing for language teachers. Cambridge: Cambridge University Press. (1989).

[11] Lumley, T. Assessment criteria in a large-scale writing test: What do they really mean to the raters? Language Testing, 19(3), 246-276. (2002).

[12] Cho, D. A study on ESL writing assessment: Intra-rater reliability of ESL compositions. Melbourne Papers in Language Testing, 8(1). (1999).

[13] Hughes, Arthur Testing for Language Teachers. Cambridge University Press, Glasgow. (1995).

[14] Bachman, F. & Alderson, J.C. Statistical analyses for language assessment. Cambridge: Cambridge University Press. (2004).

[15] Wiseman, C. S. Rater effects: Ego engagement in rater decision-making. Assessing Writing, 17, 150-173. (2012).

[16] Connors, R. & Lunsford, A. Frequency of formal errors in current college writing or Ma and Pa Kettle do research. College Composition and Communication, 39, 395-409. (1988).

[17] Lunsford, A. & Lunsford, K. 'Mistakes are a fact of life': A national comparative study. College Composition and Communication, 59, 81-806. (2008).

[18] Coffman, W. E. Essay examinations. Educational measurement, 2, 271-302. (1971).

[19] Payne, D. A. The Specification and Measurement of Learning Outcomes. Blaisdell Publishing Co., Waltham, Mass. (1968).

[20] Alderson, J.C., Clapham, C. & Wall, D. Language test construction and evaluation. Cambridge: Cambridge University Press. (1995).

[21] Weir, C.J. Language testing and validation: An evidence-based approach. New York: Palgrave Macmillan.(2005).

[22] Charney, D. The validity using holistic scales to evaluate writing: A critical overview. Research in the Teaching of English, 18(1), 65-87. (1984).

[23] Douglas, D. Understanding language testing. Chennai Micro Print (P) Ltd., India. (2011).

[24] Huot, B. The literature of direct writing assessment: Major concerns and prevailing trends. Review of Educational Research, 60, 237-263. (1990).

[25] Weigle, S.C. Effects of training on raters of ESL compositions. Language Testing, 11 (2), 197-223. (1994).

[26] Kondo-Brown, K. A FACETS analysis of rater bias in measuring Japanese second language writing performance. Language Testing, 19(1), 3-31. (2002).

[27] Weigle, S.C. Assessing writing. Cambridge: Cambridge University Press. (2002)