INTRA-RATER RELIABILITY OF HOLISTIC AND RUBRIC-BASED ASSESSMENT OF ESSAY WRITING IN PAKISTAN

Humaira Yaqub, Rabia Tabassum, Mahwish farooq
Deaprtment of Appied Linguistics
Government College University Faisalabad

ABSTRACT: *This study explores the intra-rater reliability of holistic and rubric-based assessment of essay compositions of Pakistani students. Through the variance among the results of scoring, this research proves the reliability of holistic and rubric-based assessment. The data has been collected from 200 BS students of Public Sector in Faisalabad, Pakistan. The students are selected according to convenient sampling procedure. To collect the data, the students have been given a task of writing expository composition on the essay, 'Smoking'. Afterwards, among the 200 essays, 100 essays have been selected randomly for data analysis. The data has been analyzed by one rater. The rater has scored the tests into two sessions. In the first session, the same tests have been checked holistically and with rubrics. After 10 days interval, in the second session, the same tests have been again scored holistically and with rubrics. All the results have been analyzed through SPSS software. The ANOVA analysis justifies the results in favor of rubric-based scoring. The consistency and inconsistency of scores are expressed by checking intra-rater reliability of the rater.*

Keywords: Assessment, Holistic assessment, Rubric-based assessment, Reliability,Intra-rater reliability, Rater and rating

INTRODUCTION

The present research estimates students' skills and abilities of writing a composition through scoring. Besides, reliability of assessment by the tester is focused. It is a fact that students' writings are evaluated to assess their proficiency level and to provide them a healthy feedback to develop their writing skills more for future. Writing and its assessment are actually necessary for all level courses. At all levels, there are set certain criteria to judge the writing proficiency. The criteria include assessment rules and techniques. Even rubrics of writings are used for assessment. Two types of rubrics i.e. holistic rubrics and analytic rubrics are commonly practiced. The use of rubrics is essential for assessment because many researchers claim that assessment is more reliable when teachers assess the writings with rubrics [1]. Because it never happens that rubrics have any negative effects. Moreover, the teachers may increase grading and objectivity of assessment with rubrics. It is assumed that assessment without rubrics is mere a subjective judgment [2] so it is better to use rubrics [3]. The present study reflects the concepts of rubric-based and holistic assessment. The difference between both types of assessment is analyzed. And here, the intra-rater reliability of both assessments is the matter of discussion.

REVIEW OF LITERATURE

Assessment is a process which analyzes the learning competence and knowledge of learners through their performance. McNamara [4] comments that assessment judges learner' skills. All types of assessments have certain purposes. For example, language proficiency is checked through a specific sort of assessment. Proficiency of four skills is assessed in both subjective and objective ways. Assessment is further classified into three classeses purpose wise [5]. The first purpose of is administrative which is considered a general assessment, placement, certification and promotion. The second purpose is instructional which deals with diagnostic assessment for progress, feedback and evaluation of the curriculum. The third purpose is research which includes not only knowledge of language learning but also its use. Such classification defines that there are multiple approaches with multiple options for all processes of language assessment. For these approaches the term 'Multiplism' is used [6].

Assessment is related to the language knowledge which needs to be assessed and the use of instruments or procedures for the assessment of knowledge. Assessment has importance for learners. It provides feedback regarding educational process to the learners and product to its stakeholders [7]. So, it works as two-way communication. Assessment is helpful for teachers or testers to fix the future problems and to apply necessary remedies. It also helps learners to be self-directed and to make their learning more effective [8].

Holistic Assessment

Holistic assessment involves overall judgment of learners' writing. It is opposite to analytic assessment and resembles general impression marking. Holistic assessment does follow its own criterion in testing. It is stated that general impression marking and holistic scoring are subjective [9]. A simple composition is assessed directly without setting any benchmark. No component of text is scored with fixed number. Scoring involves judgments of the tester as a whole. It is said that holistic scoring judges the predetermined components as a whole [10]. Moreover, holistic scoring is reliable to some extent. But analytic scoring achieves levels of reliability more than holistic scoring (ibid).

In holistic assessment, there are certain principles of balance, quality of assessment and involvement of students [11]. These are the features of assessment which provide guidance to institutions. They help institutions develop their assessment system and motivate learning and success of students [12]. Though holistic assessment is not much appreciated in language, yet many learning targets are achieved by it. Learners learn through feedback and assess themselves. Learning goals are set for improvement communicating the progress after holistic assessment [13]. When learners start assessing their own proficiency of language, they confidently improve themselves. Chappuis et al. [11] comments that learners' achievement is decided by the motivation which depends on the effects of assessment. The effects generated through holistic assessment are not much helpful because they may not provide step by step

guidance with subjective grading. Grading and reporting in assessment provide the educational institutions an opportunity to provoke learners for better performance because learning and development of learners is communicated by the institution after assessment [14]. It is not appropriate for the institutions to enhance learners' performance through holistic scoring.

Objections on Holistic Scoring

Without any research analysis and theoretical basis, holistic scoring is used for almost all sorts of assessment. In fact, it has some objections. Sloan and McGinnis [15] claim that holistic scoring deals with only appearance and length. Odell and Cooper [16] claim that clear assessment instructions are not given in holistic scoring so, it is useless. He further argues that holistic scoring is not true indicator of writing quality. Charney [17] claims that scoring process is different in holistic assessment, which damages the ability of raters for sound decision making. Holistic scoring actually lacks inter-rater reliability. In this reference, a study was conducted to evaluate 300 essays [18]. A nine-point scale by 53 untrained testers was used. 7 different scores were received by 94% essays. There was an infrequent agreement among the raters. The inter-rater reliability coefficient was .31 that was lower than an acceptable standard.

Rubric-based Assessment

In rubric based and analytic assessment, there is both subjectivity and objectivity. Analytic assessment is objective-like but termed as subjective systematically in which scoring criteria is applied on subjective grounds [19]. Rubrics are involved not only in holistic but also analytic assessment. There are discussed holistic rubrics which are product-oriented while analytic rubrics are process-oriented. With holistic rubrics, overall performance is assessed and analytic rubrics deals with step by step assessment to reach final result [20].

In rubric-based assessment, if there is need to assess learner's writing skill level, the components of vocabulary command, verbal ability, spelling, punctuation, and grammar are focused. According to these components, a grader evaluates the positive and negative impact of writing [21]. Rubrics guide all testers to play in their respective fields effectively. A scoring and rating criterion is selected to assess all learners on the same lines. Rubrics as tools are used to level the scoring consistency and reliability [1]. When scoring is done with rubrics, it is confidently considered reliable. Even testers are conscious and confident that they assess what rubrics demand [22]. Reliability is also increased by rubric based assessment. It has, however, been argued that due to implementation of rubrics, assessment quality, accuracy and inter-rater reliability is entertained [23].

Rubric based assessment is not always entertained positively but it is also opposed. It is noted that there is not much information regarding the effectiveness of rubrics as the tool of assessment [24]. In this assessment, raters may face some superficial and problematic factors which are not covered in rubrics, such as structure, spelling, grammar and punctuation errors [25]. Due to such reasons, language composition and writing are assessed holistically with a great reliability. Meier [26] argues that holistic scoring of composition is more reliable than analytic rubric based assessment. Actually, every kind of assessment needs different criteria. It is necessary to use rubric for every assessment. Some assessments are almost impossible to be done without defining any rubric. In fact, rubric itself is not necessary or unnecessary but its right use for the specific assessment is needed [2]. It is also stated in the favor of rubrics that assessment of all subjects can be reliably judged with rubrics [27]. Rubrics are beneficial for formative assessment. With the help of rubrics, students' work is analyzed and teachers' instructions are guided. Teachers provide the students a specific feedback through rubrics to achieve higher levels of proficiency [28] so that the students can become independent learners [29].

Assessment in Pakistan

Assessment in Pakistan has many problems. Many researchers do not favor assessment criteria applied in Pakistan. Even many researchers claim that there are no specific criteria of assessment in Pakistani institutions. It is claimed that Pakistani learners reproduce what is written in books which they memorize. The learners who is involved in sharp memorization process he gets good marks otherwise others have to face failure. It is observed that in the name of education, some books are crammed to pass the examination. Khan [30] comments this system a narrow process of evaluation. Even at primary level, educational assessment system fails to follow its mandate of learning four skills of language [31]. The focus is only to pass examination studying a selective syllabus. Passing examination with highest grades is a prestige for students, teachers and institutions. The achievement of extensive knowledge is not their motive. It is not focused whether they have met with course objectives or not. Only prestige of passing examination is necessary [32]. Rehmani [32] further defines that teaching is done for testing not for learning. If public examination and assessment system is focused specifically, there are certain problems e.g. multiple boards with lack of coordination at secondary level, conducting papers at the same time of all subjects, deceitful ways of attempting the papers and unreliable results and deficiencies in marking criteria. In this way, regular evaluation through examination does not increase the achievement level of learners but give them fatigue of cramming. Qureshi [33] identifies that there must be substantiation that regular evaluations will motivate students' achievement. It is possible by setting the criteria of assessment and providing feedback.

Moreover, it is claimed against public assessment system that some questions and selected material is given again in every year, which provide students and teachers a safe ground for assessment [34]. This is the reason of decline and ineffectiveness of assessment system because students rely on cramming without having beneficial effects to solve repeated contents of the tests and papers. So, the reliability and validity of tests are ignored [35]. In fact, there should be reliability and validity in assessment procedures. But unfortunately, testing is considered distinct from teaching and learning that causes unreliable and invalid assessment [36]. With the effect of all these problems, this study intends to identify the reliability, especially, intra-rater reliability of assessment done holistically and analytically in Pakistani context.

Reliability

Reliability involves the measurement of tests and their scores in statistical design. It is a general concept when attributes of some psychological tests need to be measured; reliability is focused [37]. Not only psychological tests but also educational tests are measured for reliability. Reliability of scores is measured applying reliability coefficient. In reliability, random measurement of errors may affect ability differences. The use of only correlation coefficient may measure reliability of tests exaggeratingly. So, some other coefficients are applied to measure reliability. Cronbach's alpha coefficient, the Kuder-Richardson 20 formula and the Spearman-Brown formula are termed as measuring tools. These tools measure valid coefficients of reliability in certain conditions [38]. Apart from these tools, for second language testing, there is applied generalizability theory and classical test theory as well. All these estimate the consistency and inconsistency of measurement. But the occurrence of consistency is necessary for exact reliability coefficients [39]. Bollen [39] further defines that reliability and its measurement should be free of errors either random or systematic. Actually in variety of conditions, reliability includes stability of measurement. Even association of measurement is one of the techniques used for reliability. The technique of standard error of measurement is used to estimates the accurate measurement in the index of reliability [40].

Types of Reliability

Reliability has its certain types e.g. test-retest reliability, alternative forms reliability, split-half tests reliability, intra-rater reliability, internal consistency reliability etc. Test-retest reliability is one of appealing forms but it has many limitations as well [37]. To check test-retest reliability, same test with a considerable interval is given twice to be solved. In this way, reliability of tests, students' performance and teachers' testing is assessed. The limitations involve that if tests are given twice with short interval, the students may solve them easily due to their memory knowledge. If tests are given twice with long interval, the students may not be able to solve them due to loss of memory knowledge or on the contrary, they may solve them better than previous one. The alternative forms reliability contains similar procedure of test-retest reliability but the tests which are conducted twice are designed with different contents. It is said that the technique of alternative forms depends on different behavior of measurement collected at alternative times [39].

The split-half tests reliability is related to the correlation between two halves of a single test. Correlation coefficient of the whole test is checked splitting the test into two equal halves [41]. It is an economic advantage that split halves are cheaper and according to practical advantage, they save time and energy. The split-half method is also used to measure variability of behaviors in the absence of alternative forms method [41].

Then inter-rater reliability depends on measuring behaviors among individuals. The internal consistency of behaviors is judged calculating their reliability coefficients [37]. Internal consistency among testers is different from consistency of tests. The reliability of internal consistency of a test is estimated correctly when there is inter-correlation among all the items of the test. To estimate reliability of item-specific variance in a uni-dimensional test, coefficient alpha is used [42]. Similarly, reliability of intra-rater consistency is also measured. This study estimates the reliability of raters who scores the same tests twice with a considerable interval.

Intra-Rater Reliability

Intra-rater reliability is one of the types of reliability. It involves repeated assessment of a test by only one rater. This study deals with such sort of reliability. Although there is not much work on intra-rater reliability, yet intra-rater reliability has been discussed in different ways. Brown, Bull, and Pendlebury [43] argue that intra-rater reliability sometimes lacks consistency. To check consistency of raters, Cronbach's alpha coefficient has been used o a great extent. Even alpha values above .70 have been reported many times and this value is considered sufficient [44]. Moreover, it is said when rubrics are used, intra-rater reliability may not be concerned. This study justifies the claim that intra-rater reliability has its concern not only with holistic but also with rubric-based assessment.

Raters and Rating

For every rating, raters paly an essential part and it is also essential for the raters that they should be trained to fix the validity of test scores [45]. It is also said that if raters are unaware of rating practices, their rating will carry no value [46]. Rater may have various reading and rating styles. A model based on experimental study consists of prototypical sequence of decision making involving three steps that raters scan the composition for surface level identification, engage in interpretation strategies and read the essay while making judgments and articulate a scoring decision summarizing judgments. This model is in favor of both experienced and inexperienced testers. The rating process designed by Lumley [47] offers that raters read and pre-score, score and revise and finalize the rating. For this process, only experienced testers are required.

Rating quality is disturbed if the raters do not apply any remarkable parameters. There are some errors of raters which affect the quality of rating [48]. Errors occur when raters have severe and lenient behaviors. Even in overall severity, raters' severity differs on the scale as some raters are more severe in rating while others are less severe [49]. Solution is that raters should be selected having equal severity of scoring. To estimate the severity and leniency level of raters, Rasch [50] approach is used. After that adjustment of scores of test takers can be done accordingly.

STATEMENT OF THE PROBLEM

In Pakistan, rubrics are not used for educational writing assessment although they are designed. The use of rubrics may not be considered reliable for assessment. To confirm that the rubrics are essential for scoring, this research investigates the effects of rubrics through intra-rater reliability. It is also assumed that holistic scoring is not more beneficial than rubric-based scoring. To test this assumption, this study focuses on consistency and inconsistency of results.

RESEARCH QUESTIONS

This study investigates research questions to draw out the findings, conclusions and implications. The following

questions have been constructed to investigate the variables of the research.

1. How consistent is the tester scoring the same test holistically after interval?
2. How consistent is the tester scoring the same test with rubrics after interval?
3. What is reliability of intra-rater assessment?

OBJECTIVES OF THE STUDY

The present research aims to analyze the reliability of two different assessments. The measurement of intra-rater reliability of holistic and analytic (rubric-based) assessment is the goal of this study. Mostly, inter-rater reliability is measured to check consistency of scores. This research provides an insight to the research how a single rater can assess essay writing in two different ways and what level of reliability their scoring reveals. It is also analyzed whether a single rater with his personal behavior of scoring makes reliability coefficients consistence or inconsistence.

SIGNIFICANCE OF THE STUDY

This study signifies the use of rubrics for assessment in Pakistani context. Rubrics are very important to be used for assessment because holistic assessments are not able to guide the students through feedback. This research actually encircles the guidelines for testers that they should use rubrics to assess the validity and reliability of their scoring. Moreover, this research also suggests the other researchers to assess the reliability of different kinds and assessments of testers. It directs the raters to use rubrics for effective assessment of students. In fact, this research provides the Pakistani raters an idea of following the very assessment that proves more reliable.

LIMITATIONS OF THE STUDY

The study encircles only intra-rater reliability of holistic and rubric-based scoring. The inter-rater reliability is not checked for two reasons. Firstly, the intra-rater reliability seems to be enough to generate the desired results. Secondly, due to space and time, the intra-rater reliability has been focused only. The analysis of results has been done through the application of ANOVA. The other types of analysis e.g. t-test, correlation coefficient etc are also appropriate to draw conclusion. The population size is not much larger in the study. It can be increase to generalize the results effectively. This study covers only one dimension of reliability of assessments. The other dimensions can also be researched.

METHODOLOGY OF RESEARCH

The research incorporates the quantitative method through which the data has been collected and analyzed. All the measurements involve in statistical empirical investigation. After collecting and analyzing the data, the results are generalized to larger population.

Data Collection and Procedure

The study has collected the data executing empirical investigation. The 200 BS students from Public Sector University in Faisalabad are selected in order to collect the data. They are given the task to write an expository composition of 500 words. The topic of the expository composition is 'Smoking'. The students are allotted an hour to complete the task of writing. The researcher with the help of three other instructors invigilates the students during their writing process. After an hour, the research collects the essays from the students to analyze them further.

Data Analysis

For the data analysis, only well-written 100 essays are selected out of 200 essays. A rater is consulted to rate the essays. The selected rater is an experienced instructor and the rater of a renowned university in Faisalabad. The instructions about the rating procedure are given to the rater beforehand. The researcher also instructs the rater about holistic and rubric-based rating. The rater is asked to rate the tests four times. He rates the tests two times holistically and two times with rubrics. At the time of data analysis, the researcher also guides the rater to avoid any inconvenience.

Time Duration

The research observes the time duration of two sessions. In the first session, the rater is directed to rate the 100 tests first holistically and then with rubrics. Two times rating of the 100 tests is done in the first session. After 10 days interval, the same rater is asked to rate the same 100 tests first holistically and then with rubrics. In fact, only one rater rates the tests four times.

Research Design

For this study, quantitative research design has been used. The description of statistical measures has been given in terms of quantitative investigation. This research describes the correlation of dependent and independent variables. The rating of the rater is a dependent variable which is tested and measured. Moreover, the explanation of the data collection, techniques of choosing a statistical procedure and the tables to provide exact values is provided in quantitative design of research. This research also presents the discussion and implications for the reliability of ratings and the consistency of the results measurements through quantitative design of research.

Population and Sample Size

For the study, the 200 BS students of Public Sector University in Faisalabad are selected as the population. All the students are enrolled in the class of BS English. Among the population of 200 students, the tests of only 100 students are chosen for the data analysis.

Sampling Technique

This study adopts the procedure of convenience sampling. All the population is conveniently selected. The students having good academic record are chosen. Even out of the 200 tests, the 100 tests are selected randomly. In fact, this research first uses convenient sampling for the selection of the population and then uses random sampling for the selection of the tests.

Instrument

This research applies the ANOVA test from SPSS 20.0 version. The intra-rater reliability of the rater is checked through the one-way completely randomized variance of two tests of holistic rating and two tests of rubric-based rating. The variance among tests justifies the consistency and inconsistency of the rating. Afterwards, the generalization of the results is applicable as a whole.

ANOVA Analysis

The ANOVA analysis is the one-way analysis of variance technique which compares the means of three or more samples. Only numerical data is displayed through this technique. A series of calculations are done in the ANOVA

analysis. It calculates the number of experimental units which summarize the treatment group for two sums, a mean and a variance. It also calculates DFs and SSs. In the ANOVA analysis, MSs are also calculated and F is determined by a ratio. Then P-value from F is also produced to know whether the results of treatments are significantly different or not. The significant results are valid. Moreover, the SS equations are simplified in the equal terms of balance experiment. In the complex experiment, extra terms of statistics are applied for analysis, which reduces the number of degrees of freedom available. Armstrong et al. [51] also describes one-way ANOVA in a randomized design to compare the reading rates of three groups of subjects including young normal subjects, elderly normal subjects and the subject with age related.

Rubrics

This study incorporates rubrics designed by University of the Punjab in Pakistan for the data analysis. The collected data has been scrutinized according to the sections of the rubric proforma (Appendix 1). All the sections of the rubric proforma have been constructed with careful attention of the testing experts. The three sections of rubrics consist of grammar, vocabulary, punctuation and the structure of English language. The checking of the quality of language is also the part of rubrics. All the points in the rubric proforma cover almost everything which is checked in English compositions at BS level and no point is superfluous.

FINDINGS AND DISCUSSION

The findings indicate the difference between the results of ratings. First of all, the tables display the student IDs with their scores. These scores are further analyzed through ANOVA to draw out their variance. The results of scoring and the variance of the scores have been displayed as well.

Table.1 The Description of Holistic and Rubric-Based Scoring

HOLISTIC SCORING					
Student-IDs	**1st Session**	**2nd**	**Student-IDs**	**1st Session**	**2nd Session**
BS-A-01	14	10	**BS-B-01**	19	21
BS-A-02	12	11	**BS-B-02**	18	20
BS-A-03	12	12	**BS-B-03**	18	19
BS-A-04	15	16	**BS-B-04**	20	18
BS-A-05	16	17	**BS-B-05**	20	19
BS-A-06	18	19	**BS-B-06**	22	22
BS-A-07	18	19	**BS-B-07**	21	20
BS-A-08	10	15	**BS-B-08**	15	16
BS-A-09	12	13	**BS-B-09**	18	16
BS-A-10	16	10	**BS-B-10**	20	21
BS-A-11	12	12	**BS-B-11**	17	16
BS-A-12	16	14	**BS-B-12**	19	21
BS-A-13	20	16	**BS-B-13**	25	25
BS-A-14	12	13	**BS-B-14**	18	17
BS-A-15	15	19	**BS-B-15**	19	19
BS-A-16	9	6	**BS-B-16**	12	13
BS-A-17	9	8	**BS-B-17**	11	11
BS-A-18	9	9	**BS-B-18**	13	12
BS-A-19	10	10	**BS-B-19**	15	12
BS-A-20	8	10	**BS-B-20**	12	14
BS-A-21	10	8	**BS-B-21**	14	15
BS-A-22	15	14	**BS-B-22**	19	18
BS-A-23	12	9	**BS-B-23**	19	20
BS-A-24	11	6	**BS-B-24**	16	16
BS-A-25	11	7	**BS-B-25**	15	14
BS-A-26	19	18	**BS-B-26**	22	22
BS-A-27	17	13	**BS-B-27**	20	19
BS-A-28	18	12	**BS-B-28**	22	22
BS-A-29	17	11	**BS-B-29**	19	18
BS-A-30	19	15	**BS-B-30**	23	23
BS-A-31	15	11	**BS-B-31**	20	20

BS-A-32	16	16	**BS-B-32**	19	17
BS-A-33	12	9	**BS-B-33**	13	12
BS-A-34	13	8	**BS-B-34**	14	14
BS-A-35	10	7	**BS-B-35**	14	14
BS-A-36	16	12	**BS-B-36**	18	18
BS-A-37	9	10	**BS-B-37**	11	13
BS-A-38	8	8	**BS-B-38**	12	12
BS-A-39	22	20	**BS-B-39**	25	24
BS-A-40	12	12	**BS-B-40**	15	16
BS-A-41	15	10	**BS-B-41**	17	20
BS-A-42	13	15	**BS-B-42**	16	16
BS-A-43	16	13	**BS-B-43**	20	21
BS-A-44	18	16	**BS-B-44**	21	21
BS-A-45	9	9	**BS-B-45**	12	13
BS-A-46	7	11	**BS-B-46**	12	12
BS-A-47	14	14	**BS-B-47**	19	18
BS-A-48	11	13	**BS-B-48**	14	15
BS-A-49	12	12	**BS-B-49**	18	16
BS-A-50	16	17	**BS-B-50**	20	20
BS-A-51	8	7	**BS-B-51**	13	14
BS-A-52	18	14	**BS-B-52**	22	23
BS-A-53	17	12	**BS-B-53**	21	21
BS-A-54	20	22	**BS-B-54**	24	25
BS-A-55	12	13	**BS-B-55**	19	19
BS-A-56	11	9	**BS-B-56**	17	18
BS-A-57	9	12	**BS-B-57**	15	15
BS-A-58	17	15	**BS-B-58**	20	20
BS-A-59	12	14	**BS-B-59**	19	18
BS-A-60	6	5	**BS-B-60**	15	15
BS-A-61	19	17	**BS-B-61**	23	23
BS-A-62	19	18	**BS-B-62**	22	21
BS-A-63	12	11	**BS-B-63**	19	19
BS-A-64	5	8	**BS-B-64**	14	13
BS-A-65	9	13	**BS-B-65**	16	16
BS-A-66	18	19	**BS-B-66**	22	22
BS-A-67	10	12	**BS-B-67**	17	16
BS-A-68	3	5	**BS-B-68**	12	13
BS-A-69	5	6	**BS-B-69**	13	13
BS-A-70	15	21	**BS-B-70**	22	22
BS-A-71	8	9	**BS-B-71**	14	13
BS-A-72	11	11	**BS-B-72**	19	17
BS-A-73	6	10	**BS-B-73**	16	16
BS-A-74	6	9	**BS-B-74**	13	13
BS-A-75	10	12	**BS-B-75**	18	19
BS-A-76	5	8	**BS-B-76**	15	16
BS-A-77	11	12	**BS-B-77**	19	19
BS-A-78	19	19	**BS-B-78**	23	23
BS-A-79	9	10	**BS-B-79**	16	17

BS-A-80	18	14	**BS-B-80**	21	21
BS-A-81	12	7	**BS-B-81**	18	16
BS-A-82	18	19	**BS-B-82**	22	21
BS-A-83	17	16	**BS-B-83**	23	23
BS-A-84	19	19	**BS-B-84**	24	25
BS-A-85	3	3	**BS-B-85**	11	11
BS-A-86	22	20	**BS-B-86**	25	25
BS-A-87	14	9	**BS-B-87**	19	19
BS-A-88	19	20	**BS-B-88**	24	23
BS-A-89	8	11	**BS-B-89**	18	18
BS-A-90	16	16	**BS-B-90**	20	21
BS-A-91	13	12	**BS-B-91**	17	17
BS-A-92	15	15	**BS-B-92**	20	20
BS-A-93	2	3	**BS-B-93**	12	13
BS-A-94	11	13	**BS-B-94**	18	18
BS-A-95	12	8	**BS-B-95**	17	17
BS-A-96	11	15	**BS-B-96**	20	20
BS-A-97	8	4	**BS-B-97**	13	11
BS-A-98	20	18	**BS-B-98**	23	23
BS-A-99	16	12	**BS-B-99**	20	18
BS-A-100	17	20	**BS-B-100**	23	23
Total	1297	1242	**Total**	1797	1793

Table (1) arranges the scores of holistic and rubric-based ratings. The tests have been given IDs for their organization. In the table, the scores are organized according to students IDs. Total scores of holistic rating in the 1st session are higher than the scores in the 2nd session. The tests in the both sessions are scored by the same rater but the rater's scoring is calculated with differences. Such difference restricts the scores to be reliable. Rubric-based rating discusses the total scores in the 1st session and the 2nd session with the least difference of the rater's rating. Here, it is noticed that holistic scores of two sessions are less reliable than the rubric-based scores of two sessions. In fact, the reliability of scores is obvious in rubric-based rating. The further analysis of the scores through ANOVA test manifests the results with variance.

Analysis of Results of Holistic and Rubric-Based Scoring

The results of holistic and rubric-based scoring are obvious. The sum of two tests of holistic scoring and the sum of two tests of rubric-based scoring expose a remarkable difference in the table (1). Viewing the sum of scores, it is judged that rubric-based scoring is reliable because it shows little difference of scores. The scores are reliable due to their reliable coefficients which do not show a greater inconsistency of results. Lado [52] reports that the considerable reliability coefficient is 0.90-0.99. The following table exhibits the reliable co-efficiency of scores in rubric-based scoring.

Table.2 The Results of Holistic and Rubric-Based Scoring

Holistic Scoring	**1st Session**	**2nd Session**	**Rubric-based**	**1st Session**	**2nd Session**
n	100	100	**n**	100	100
$\bar{X}$	12.970	12.420	**$\bar{X}$**	17.970	17.930
s	4.543	4.360	**s**	3.713	3.718
Xave	12.695		**Xave**	17.950	
Variance	20.34	18.8236	**Variance**	13.6491	13.6851

Table (2) exhibits the difference among the values of *n*= sample size, $\bar{X}$= sample mean, *s*= sample standard deviation and *Xave*= average. The variance of the four test groups is also demonstrated in the table. In the four tests groups, the sample size *n*= 100 is equal. The sample mean $\bar{X}$= 12.970 is calculated in the 1st session of holistic scoring and the sample mean $\bar{X}$= 12.420 is calculated in the 2nd session of holistic scoring. Both the sessions indicate the higher difference between the values of mean and are determined less reliable. The calculated sample mean $\bar{X}$= 17.970 in the 1st session shows the less distinction than the calculated sample mean $\bar{X}$= 17.930 in the 2nd session of rubric-based scoring. Here,

both the sessions indicate the lower difference between values of mean and are determined more reliable. Same is the case with the values of standard deviation. In the two sessions of holistic scoring, the values of standard deviation $\underline{s}$= 4.543 and $\underline{s}$= 4.360 do not have reliability due to higher difference. In the two sessions of rubric-based scoring, the value of standard deviation $\underline{s}$= 3.713 and $\underline{s}$= 3.718 have reliability due to lower difference. The variance of 20.34 and 18.8236 in two holistic scoring and the variance of 13.6491 and 13.6851 in two rubric-based scoring justify the reliability of results in rubric-based scoring.

Table.3 One-way Completely Randomized Analysis of Holistic Scoring

Source	df	SS	MS	F	P-value
Treatments	1	15.125	15.125	0.7629	0.7474
Error	198	3925.270	19.825		
Total	199	3940.395			

ANOVA Analysis of Results

The ANOVA analysis of the results also describes the reliability of rubric-based scoring. The application of ANOVA analyzes the sources of variation including treatments, error and total values. It determines the values according to degree of freedom, sum of squares, mean squares, F-ratio and P-value. The subsequent table defines the values on the results of holistic scoring.

Table (3) includes the degrees of freedom in terms of treatments, errors and total. The sum of squares, mean squares and P-value are also analyzed in the table. The goal of any statistical analysis is to know the link between variables and observations. The ANOVA table (3) describes the values following the same statistics. Here the value on degrees of freedom is the function of sample size [53]. The sample size n= 100 of this research also connects with the degrees of freedom. The table shows the holistic scoring differences with statistical measurements. The following table follows the same criterion.

Table 4. One-way Completely Randomized Analysis of Rubric-Based Scoring

Source	df	SS	MS	F	P-value
Treatments	1	0.080	0.080	0.0058	0.9977
Error	198	2733.420	13.805		
Total	199	2733.500			

Both tables (3) and (4) present the output of ANOVA analysis. The degrees of freedom in both holistic and rubric-based scoring depict equal statistics. The sample size of four tests groups in both types of scoring is equal. The statistical measurements of sum of squares and mean square are different according to different scores of tests. Total variation of sum of squares and mean squares is different in both types of scoring. The P-value in the analysis of holistic scoring is less than the P-value in the analysis of rubric-based scoring. The comparison of tables (3) and (4) justifies the reliability of rubric-based scoring with significant statistics.

Graphics of ANOVA Analysis

The next graphs display the outcomes of ANOVA analysis. Two separate graphs have been drawn for the description of holistic and rubric-based scoring.

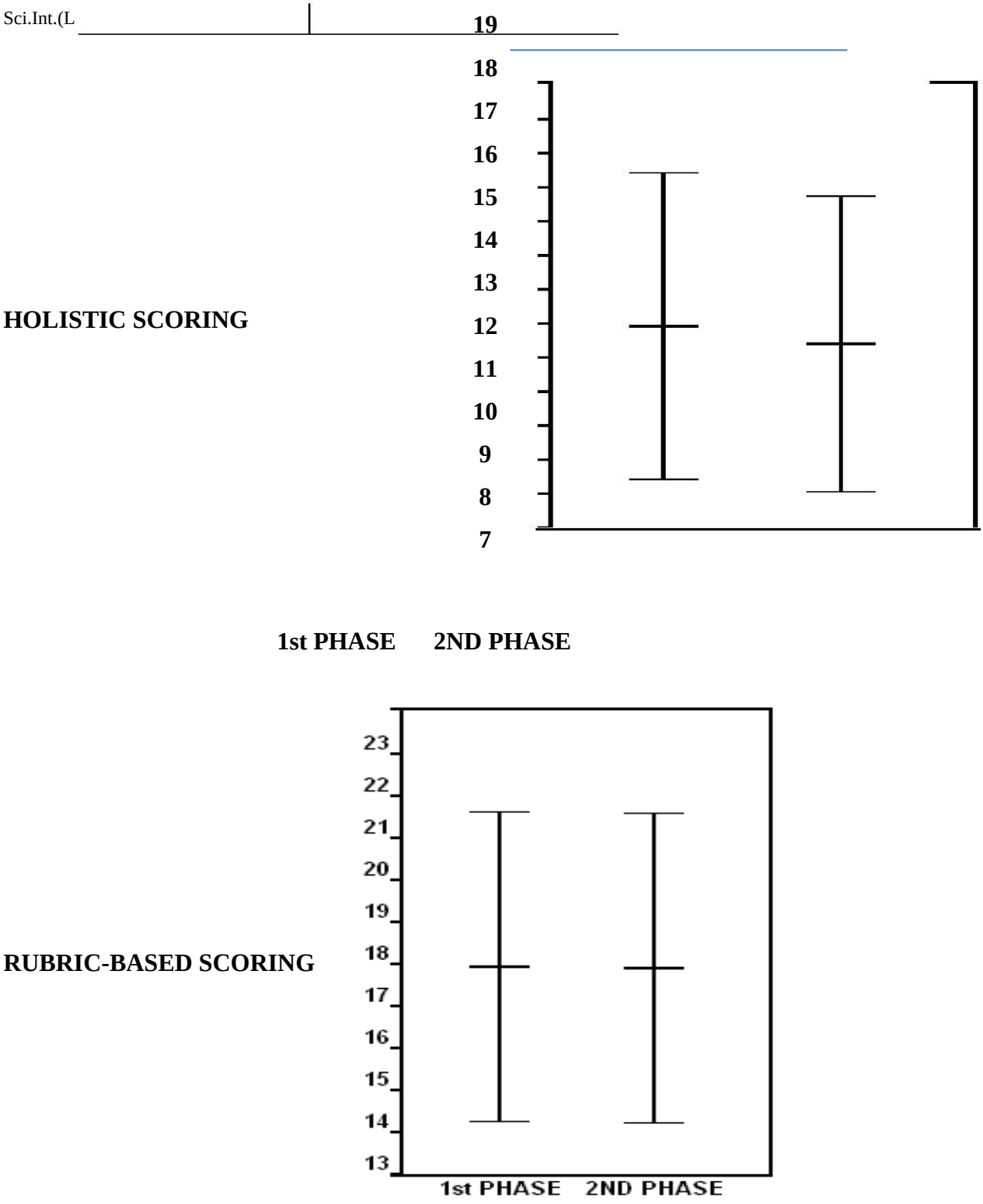
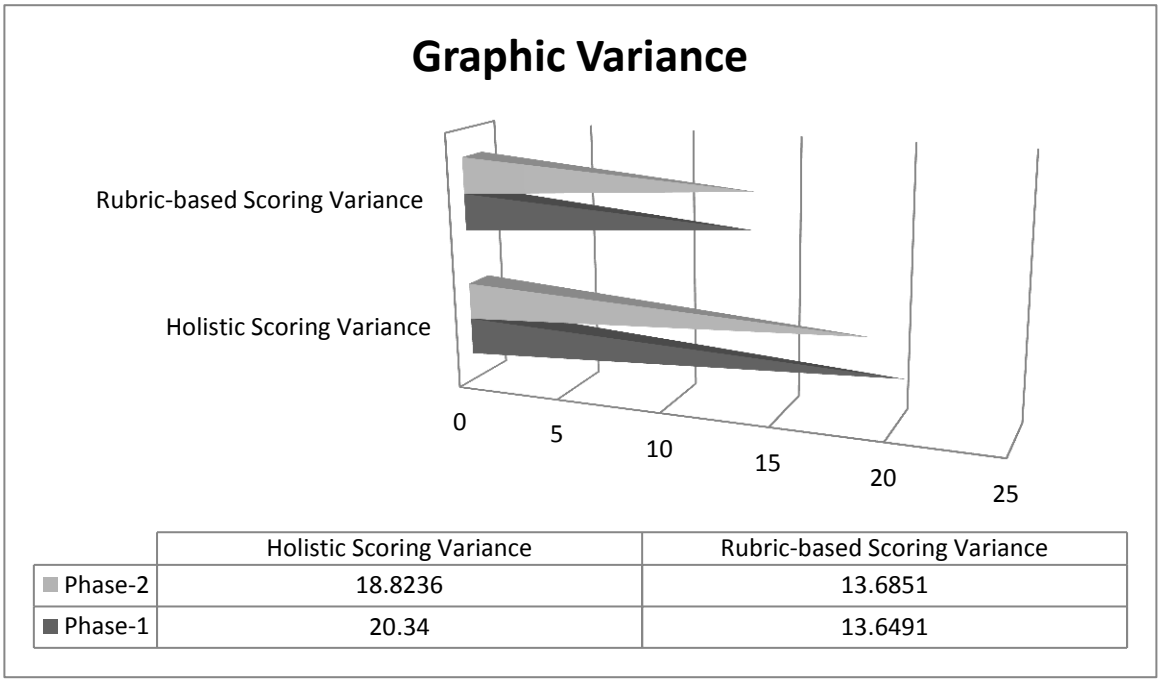


Figure. 1 Graphic Variance of ANOVA Analysis of Holistic and Rubric-based Scoring

Figure (1) shows the graphs of holistic and rubric-based scoring results. The graph showing the results of holistic scoring contains two lines having no proportion or sequence. The un-sequential lines represent the results that holistic scoring receives inconsistencies between two results. The holistic scoring before interval and after interval is not consistent. The graph of rubric-based scoring defines two results in a sequence. The results are consistent between two scoring. The interval does not affect the consistency of scoring results. So, the reliability of rubric-based scoring is justified through the graphical description. The succeeding figure is also drawn to confirm the reliability of rubric-based scoring.



	Holistic Scoring Variance	Rubric-based Scoring Variance
Phase-2	18.8236	13.6851
Phase-1	20.34	13.6491

Figure.2 Graphic Variance of Holistic and Rubric-based Analysis

Figure (2) demonstrates the variance between the results of holistic and rubric-based scoring. The variance between the phase (1) and the phase (2) of holistic scoring is more than the variance between the phase (1) and phase (2) of rubric-based scoring. In fact, the results justify the reliability of rubric-based scoring.

CONCLUSIONS

The study justifies the reliability of intra-rater assessment with rubrics and without rubrics. Intra-rater assessment with rubrics is more reliable than without rubrics because intra-rater assessment, done with rubrics, gives similar results after even a number of intervals. But intra-rater assessment, done holistically, does not give similar results after a number of intervals. So, the reliability of rubric-based scoring is confirmed. In holistic assessment, the scoring of testers is influenced by many factors and the results on reliability are affected. This study investigates the three questions. The first question inquires the consistency of the tester scoring the same test holistically with intervals. The investigation of this question clarifies that holistic scoring is unable to show reliability because its scores are inconsistent. In this research, holistic scoring has been done twice by the same rater. The scores indicate greater difference and inconsistency. Due to the inconsistent scores, the holistic scoring is not claimed a reliable method of assessment.

The second question is related to the investigation of the reliability of rubric-based scoring. The rubric-based scoring has also been done twice by the same rater. The rater's scores indicate the consistency even after interval. Such consistency determines the reliability of rubric-based scoring and justifies it a reliable method of assessment. The third question is about the reliability of intra-rater assessment. It is also justified that intra-rater assessment is reliable in terms of consistency of results. If the rater scores the tests even more than two times, rubric-based scoring gives the similar and reliable results. This research achieves the goal of determining the importance of rubric-based assessment. Though holistic assessment is always not neglected, yet it is better to choose rubric-based assessments to draw convenient outcomes. This research also explains the solution to the problem that rubric-based scoring is not appreciated for many assessments in general and for the assessment in Pakistani context in particular. The results of this research provide awareness to the raters that rubric-based scoring is more reliable than holistic scoring. In fact, this study convinces all the testers in general and the Pakistani testers in particular to use rubrics for effective assessment.

RECOMMENDATIONS

This study recommends the application of rubric-based scoring for Pakistani assessments. It also recommends the researchers to research other dimensions which can convince the raters to score the tests with the help of rubrics. The findings of this research suggest the raters to design the rubrics for assessment purposes so that their assessment could be a useful feedback for students. Such study can be conducted further by designing rubrics for all language skills. Following the format of rubrics given in this study, the rubrics to test other language skills can easily be constructed. The trend of holistic scoring needs to be followed when it is essential in some cases. But rubric-based scoring proves useful for all sorts of assessments. Moreover, this study has been conducted on micro level. It is recommended that a macro level study can be conducted by increasing the population size and the number of raters. Due to increase in population size and the number of raters, the better results can be generated. The study with a large population can better determine the reliability of rubric-based assessment.

REFERENCES

[1] Jonsson, A., & Svingby, G. (2007). The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences. Educational Research Review, 2, 130–144.

[2] Turley, E. D., & Gallagher, C. G. (2008). On the Uses of Rubrics: Reframing the Great Rubric Debate. English Journal, 79 (4), 87–92.

[3] Spandel, V. (2006). In Defense of Rubrics. English Journal, 96 (1), 19–22.

[4] McNamara, T. F. (1996). Measuring Second Language Performance. Essex: Longman.

[5] Shepard, L. A. (2000). The Role of Assessment in a Learning Culture. Paper presented at the Annual Meeting of the American Educational Research Association. Available http://www.aera.net/meeting/am2000/wrap/praddr01.html.

[6] Shohamy, E. (1998). Critical Language Testing and Beyond. *Studies in Educational Evaluation*. 24, 4, 331-345.

[7] McAlpine, M. (2002). Principles of Assessment. CAA Centre, University of Luton, www.caacentre.ac.uk/resources/bluepapers/index.shtml, last accessed 28 October 2007.

[8] Angelo, T. A., & Cross, K. P. (1993). Classroom Assessment Techniques: A Handbook for College Teachers, Second Edition, San Francisco: Jossey-Bass Publishers.

[9] Vaughan C. (1991). Holistic Assessment: What goes on in the Rater's Mind? L. Hamp Lyons (Ed.), In Assessing Second Language Writing in Academic Contexts (p. 111- 126). Norwood, NJ: Ablex.

[10] Johnson, R., Penny, J., & Gordon, B. (2001). Score Resolution and the Inter-Rater Reliability of Holistic Scores in Rating Essays. *Written Communication, 18*(2), 229- 249.

[11] Chappuis, S., Commodore, C., & Stiggins, R. (2010). Assessment Balance and Quality: An Action Guide for School Leaders. Portland, OR: Assessment Training Institute.

[12] Hattie, J. A. C. (2009). Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement. London, UK: Routledge.

[13] Stiggins, R. J. (2000). Classroom Assessment: A History of Neglect, a Future of Immense Potential. Paper presented at the Annual Meeting of the American Educational Research Association.

[14] Guskey, T. R., & Bailey, J. M. (2010). Developing Standards-Based Report Cards. Thousand Oaks, CA: Corwin.

[15] Sloan, C. A., & McGinnis, I. (1978). The Effects of Handwriting on Teachers' Grading of High School Essays. ERIC, 1978. ED 220.

[16] Odell, L., & Cooper, C. R. (1980). Procedures for Evaluating Writing: Assumptions and Research. College English 42 (Sept. 1980): 35.

[17] Charney, D. A. (1984). The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview. Research in the Teaching of English 18 (Feb. 1984): 65-81.

[18] Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in the Judgment of Writing Quality. Princeton: Educational Testing Service, 1961. ETS RB.

[19] Kayapinar, U. (2010). A Study on Assessment Tools and Evaluation of Essay Writing Skill in Foreign Language Education. Unpublished PhD Dissertation, Mersin University. Yenisehir Campus: Turkey.

[20] Finson, K. D. (1998). Rubrics and their Use in Inclusive Science. Intervention in School and Clinic, 34 (2), 79–88.

[21] Read, B., Francis, B., & Robson, J. (2005). Gender, Bias, Assessment and Feedback: Analyzing the Written Assessment of Undergraduate History Essays. Assessment and Evaluation in Higher Education, 30 (3), 241–260.

[22] Silvestri, L., & Oescher, J. (2006). Using Rubrics to Increase the Reliability of Assessment in Health Classes. International Electronic Journal of Health Education, 9, 25–30.

[23] Kohn, A. (2006). The Trouble with Rubrics. English Journal, 95 (4), 12–15.

[24] Hafner, J. C., & Hafner, P. M. (2003). Quantitative Analysis of the Rubric as an Assessment Tool: An Empirical Study of Student Peer-Group Rating. International Journal of Science Education, 25 (12), 1509–1528.

[25] Elliot, N. (2005). On a Scale: A Social History of Writing Assessment in America. New York: Peter Lang.

[26] Meier, S. L. (2006). Teachers' Use of Rubrics to Score Non-Traditional Tasks: Factors Related to Discrepancies in Scoring. Assessment in Education: Principles, Policy and Practice, 13 (1), 69–95.

[27] Brookhart, S. M. (2005). The Quality of Local District Assessments Usedin Nebraska's School Based Teacher-led Assessment and Reporting System (STARS). Educational Measurement: Issues and Practice, 24 (2), 14–21.

[28] Bush, W. S., & Leinwand, S. (2000). eds. Mathematics Assessment: A Practical Handbook for Grades 6-8. Reston, VA: NCTM.

[29] Guskey, T. R. (2003). How Classroom Assessments Improve Learning. Educational Leadership 60, no. 5 (2003): 6-11.

[30] Khan, S. (2006). An Evaluation of the Exercises Provided in the English Compulsory Textbook for Class X, [Unpublished MA dissertation] Faculty of English Linguistics, University of Karachi.

[31] UNESCO (United Nation Educational Scientific and Cultural Organization), (2007). The Education System in Pakistan: Assessment of the National Education Census, Islamabad: UNESCO.

[32] Rehmani, A. (2003). Impact of Public Examination System on Teaching and Learning in Pakistan. International Biannual Newsletter ANTRIEP, 8 (2) Pp.3-7.

[33] Qureshi, Q. B. (2005). Clinical Assessment of Children with Disabilities Child: Care. Health and Development, 31(4) 497.

[34] Christie, T., & Afzaal, M. (2005). Rote Memorization as a Sufficient Explanation of Secondary School Examination Achievement in Pakistan: An Empirical Investigation of a Widespread Assumption., Paper Presented in the conference Assessment and the future schooling and learning held in Abuja, Nigeria. Retrieved from: http://www.aku.edu/AKUEB/pdfs/IAEA05. pdf (Accessed: 27/12/2011).

[35] Khan, I. (2011). Reading Assessment Techniques among Selected Secondary School Teachers in Pakistan: Current Trends and Practices, International Journal on New Trends in Education and their Implications, 2 (2) Pp.58-75.

[36] Ahmed, S., & Malik, S. (2011). Examination Scheme at Secondary School Level in Pakistan: Composite vs Split, Canadian Social Science 7 (1) Pp. 130-139.

[37] Rosenthal, R., & Rosnow, R. L. (1991). Essentials of Behavioral Research: Methods and Data Analysis. Second Edition. McGraw-Hill Publishing Company, pp. 46-65.

[38] Zimmerman, D.W. (1972). Test Reliability and the Kuder-Richardson Formulas: Derivation from Probabilistic Theory. Educational and Psychological Measurement 32, 939-54.

[39] Bollen, K. A. (1989). Structural Equations with Latent Variables (pp. 179-225). John Wiley & Sons.

[40] Feldt, L. S., & Qualls, A. L. (1999). Variability in Reliability Coefficients and the Standard Error of Measurement from School District to District. Applied Measurement in Education, 12 (4), 367–381.

[41] Nunnally, J. C. (1978). Psychometric Theory. McGraw-Hill Book Company, pp. 86-113, 190 255.

[42] Cortina, J. M. (1993). What is Coefficient Alpha? An Examination of Theory and Applications. Journal of Applied Psychology, 78 (1), 98-104.

[43] Brown, G., Bull, J., & Pendlebury, M. (1997). Assessing Student Learning in Higher Education. London: Routledge.

[44] Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the Scoring of Writing: Studies of Reliability and Validity Using a New Zealand Writing Assessment System. Assessing Writing, 9, 105–121.

[45] Lumley, T. (2002). Assessment Criteria in a Large-Scale Writing Test: What do They really Mean to the Raters?, Language Testing, 19, 246-276.

[46] Connor-Linton, J. (1995a). Looking Behind the Curtain: What do L2 Composition Ratings Really Mean? TESOL Quarterly, 29, pp. 762-765.

[47] Lumley, T. (2006). Assessing Second Language Writing: The Rater's Perspective. Frankfurt am Main: Peter Lang.

[48] Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the Ratings: Assessing the Psychometric Quality of Rating Data. Psychological Bulletin, 88(2), pp. 413-428.

[49] Schaefer, E. (2008). Rater Bias Patterns in an EFL Writing Assessment. Language Testing, 25(4), 465-493.

[50] Rasch, G. (1980). Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: University of Chicago Press.

[51] Armstrong, R. A., Slade, S. V. & Eperjesi, F. (2000). An Introduction to Analysis of Variance (ANOVA) with Special Reference to Data from Clinical Experiments in Optometry. *Ophthal Physiol. Opt 20*, 235–241.

[52] Lado, R. (1961). *Language Testing*. London: Longman.

[53] Trochim, W. M. K. (2005). Research Methods: The Concise Knowledge Base. Cincinnati: Atomic Dog.

APPENDIX-01

CRITERIA		ATTRIBUTES	4	3	2	1	0
ORGANIZATION	A.1. INTRODUCTION						
	A.1.1. Introductory Sentences	Effective introductory sentences					
	A.1.2. Thesis Statement	Appropriate thesis statement (thesis and central idea)					
	A.2. BODY PARAGRAPHS						
	A.2.1. Topic Sentence	Appropriate topic sentence (possibly implied) supporting the thesis and the central idea					
	A.2.2. Supporting Sentences	Appropriate sentences supporting the topic (possibly major and minor)					
	A.3. CONCLUSION	Appropriate conclusion related to thesis					
LANGUAGE USE	B.1. Word Order	Correct word order					
	B.2. Pattern Variety	Using different patterns					
	B.3. Verb Form	Using verb forms correctly					
	B.4. Tenses	Using tenses appropriately					
	B.5. Articles	Using articles correctly					
	B.6. Pronouns	Using pronouns correctly					
	B.7. Prepositions	Using prepositions correctly (verb + preposition, adjective + preposition)					
VOCABULARY	C.1. Word Choice	Selecting the appropriate words					
	C.2. Word Variety	Having a rich vocabulary					