British International Journal of Education And Social Sciences
Vol.7, No.12; December- 2020
ISSN (3342 – 543X);
p –ISSN (4519 – 6511)
Impact factor: 8.11

INFUENCE OFANALYTC AND HOLISTIC SCORING PATTERNSON SCOREER RELIABILITYOF GEOGRAPHY ESSAY TEST IN OBOLLO AFOR EDUCATION ZONE OF ENUGU STATE

Charles Okafor Ugwuanyi
Department of Science and Vocational Education, Godfrey Okoye University, Enugu.

Abstract: The continuous low scorer reliability involved in scoring of essay tests is a cause for worry to everyone that is involved in assessment processes. The glaring inconsistencies have been largely due to patterns/methods adopted during scoring exercises. In view of the above problems, the study was designed to investigate the influence of analytic and holistic scoring patterns on scorer reliability of geography essay test in Obollo Afor Education Zone of Enugu State. Ex – post facto research design was used. All the 47 geography teachers and 337 SS3 students from 46 public secondary schools in Obollo Afor Education Zone were used. The instrument for data collection was a sample of Geography Essay Test (GET) that was adopted by the researcher from West African Senior Secondary Certificate Examination (WASSCE) of November, 1999. The mean and standard deviation were used to obtain the associate descriptive statistics while Spearman's Rank Order Correlation Coefficient was used to obtain the reliability estimates within groups. The results showed that the use of holistic scoring pattern in the assessment of GET has higher correlation coefficient and more reliable followed by analytic scoring pattern in the same test. Based on the findings, it was recommended that holistic scoring pattern be adopted for use in Nigerian secondary schools.

Keywords: Analytic And Holistic Scoring, Patternson Scoreer Reliability, Geography Essay Test, Obollo Afor Education Zone

Introduction

The assessment of essay tests is a complex activity that is subject to human judgment. Accordingly, this makes it challenging to achieve a fair, accurate and reliable assessment of students writing. Scoring patterns are those methods that are employed in the scoring of essay tests in order to achieve higher scorer reliability. Some of the scoring patterns are conventional patterns of scoring all items in a script; ranking all scripts before scoring all items, re – arrangement of the order of the papers before scoring, dividing the task of scoring into sessions, scoring an item across board, analytic and holistic scoring patterns (Ebuoh, 2018). However, scoring patterns have been substantially classified as either holistic: where a single score is given to each writing sample; or analytic: which separate scores given to different aspects of writing such as content, organization, language use and so on (Gonzalez, Trejo and Roux, 2017). This glaring inconsistency becomes manifest in the assessment of geography essay test. Geography is fundamentally an interdisciplinary subject which studies places, space and the environment (Institute of Australian geographers, 2020). Geography is often defined in terms of its branches: physical and human geography. The physical geography deals with the study of processes and patterns in the natural environment such as the atmosphere, hydrosphere,

British International Journal of Education And Social Sciences
Vol.7, No.12; December- 2020
ISSN (3342 – 543X);
p –ISSN (4519 – 6511)
Impact factor: 8.11

biosphere and geosphere, while human geography deals with the study of people and their localities, cultures, economies and their interactions with the environment (Pidwirny, 2016).

Since students learning and assessment in geography are ongoing processes involving the systematic collection, examination, interpretation and use of evidence to document and improve student learning, scoring patterns that articulate expectations of assessments is a key aspect of developing effective measures(Akintade, 2011). Scoring of essay tests had been criticized for not being reliable because of the use of inadequate scoring patterns. There was evidence to show that the level of unreliability of scores appears more when both holistic and analytic patterns are used interchangeably than if only one of the patterns is applied.

This is more apparent if we consider the fact that despite the standardized nature of a senior secondary school certificate geography examination of 1999 WASSCE with its marking guide a significant variations were found in the scores. This is exemplified by the result of the geography essay scores by two groups of raters (Table 1).

Table 1.Scores Given by Holistic and Analytic Scorers

Scorers	1st *Rater*	2nd *Rater*	3rd *Rater*
Holistic Scores	68	72	70
Analytic Scores	55	63	58

Scores given by 3 raters each in the script in scoring Geography essay test.

Table 1 shows that out of a randomly selected script, each of the scorers involved in scoring such script gave different marks. These differences occurred among the different scorers who applied holistic scoring pattern and analytic scoring pattern. This is more worrisome even as both scorers were served with the respective scoring guides. Further analysis of the relationship between the Holistic scorers using Spearman's Rank Order Statistic indicated a coefficient value of 0.763. Those that applied Analytic scoring pattern had coefficient value of 0.699. Expectedly, scorers with scoring/marking guides and using the same scoring pattern would have had a very strong correlation coefficient. It was claimed that low level of scorer reliability has patterns in scoring essay tests such as analytic and holistic scoring patterns.

In analytic scoring, the teacher adopts a list of the major points he expects the students to include in the response(Brown et al,2014). The analytic scoring pattern is more suitable for scoring the restricted response format. For example; "State three features of a computer" (3 marks).

Answer Key: 1 mark each for any of the following points:

a. Storage capacity.
b. Versatility.
c. Reliability.
d. Durability.
e. Programmable.
f. Fast or speed.

Here, the testee is restricted to just mentioning any three of the features of computer and so on. It is imperative to note that whether restricted or otherwise analytic however provides some merits and demerits in the scoring pattern:

Advantages of the Analytic Scoring Pattern

1. It helps teachers to keep the full range of writing features in mind as they score.
2. The diagnostic nature of analytic scoring helps students to know areas they need to improve on in the essay writing skill.
3. It helps to minimize the Halo effects and leniency error. Halo effect in the tendency for an impression in one area to influence opinion in another area. While leniency error occurs when a teacher rates or scores students too positively.

British International Journal of Education And Social Sciences
Vol.7, No.12; December- 2020
ISSN (3342 – 543X);
p –ISSN (4519 – 6511)
Impact factor: 8.11

4. It is very good for scoring large number of short specific items.

Disadvantages of Analytic Scoring Pattern

1. It is time consuming. For each piece of writing, the teacher is expected to make not less than eleven separate judgments.
2. Students most times do not make use of the analytic comments (feedback) from the teachers.
3. Negative feedback can be psychologically destructive.

A Guide to Maximize Effective Analytic Scoring

a. Analytic scale should be designed in such a way that it will help to define grading criteria clearly. It should also be shared with the students to help them understand what is expected of them and how their responses would be assessed.
b. Criteria are weighed according to their importance. For instance, if the goal of a test is to ascertain the level of assimilation of course material, then logic, ideas, organizational skill and ingenuity are scored higher than grammar and mechanics.
c. Feedback becomes formative and effective when the comments are balanced and both support challenged students.
d. Teachers should as much as possible avoid sarcasm in their comments, cancelling student's work with lines and all other forms of destructive criticism (Ebuoh, 2018).

2. Holistic Scoring Pattern

This involves the teachers selecting some students' answers that are graded as high, average or low achievements. These selected answers become the model by which the teacher assesses and scores the other answer scripts. This style of scoring is most appropriate for scoring long unrestricted essay (Weighle, 2002).

Advantages of Holistic Scoring Pattern

1. It is more reliable than analytic scoring because it requires that the teacher and one or more readers/scorers read the essays to determine which is stronger or weaker among the model essays.
2. It is efficient and takes less time than analytic scoring.
3. It saves time by minimizing the number of decisions raters make. It can be applied consistently by trained raters, hence, increasing reliability.

Disadvantages of Holistic Scoring Pattern

1. The model essay cannot be given to students for comparison. Unlike the analytic scoring where the teacher includes formative comments, the holistic scoring does not make this available for students.
2. Holistic scoring is impracticable for individual use. It is better used as a team.
3. It does not provide specific feedback for improvement: When student work is at varying levels spanning the criteria points, it can be difficult to select single best description and criteria cannot be weighted.

The above characteristics depict the nature of the scoring patterns this work intends to investigate its reliability. The main purpose of the study was to find out the influence of analytic and holistic scoring patterns on scorer reliability to geography essay test.

The research questions were formulated to guide the study:

- What is the influence of analytic scoring pattern on scorer reliability in Geography essay test?
- What is the influence of Holistic scoring pattern on scorer reliability in Geography essay test?

The Null Hypotheses were tested at 0.05 level of significance:

- Holistic scoring pattern does not significantly influence scorer reliability in geography essay test.
- Analytic scoring pattern does not significantly influence scorer reliability in geography essay test.

3 Methods

The study adopted "ex-post facto design". The subjects were randomly selected and assigned to groups by randomization. No pre – test was used. The randomization controls for the possible extraneous variables and assures that any initial difference between groups is due only to chance and followed the laws of probability (Nworgu, 2015). The scorers were randomly assigned to

British International Journal of Education And Social Sciences
Vol.7, No.12; December- 2020
ISSN (3342 – 543X);
p –ISSN (4519 – 6511)
Impact factor: 8.11

experimental group I, and experimental group II. The treatment of the subject (scores) was done and indicated below:

Table 1: Assignment of Scorers to Treatment Groups:

	Randomization	Groups	Independent Variables	Post-test Scores
Experimental Group I	R		ASP	
	R	E1	HSP	02
Experimental Group II		E2		02

Where:

E_1 = Experimental Group One

E_2 = Experimental Group Two

0_2 = Post-test Treatment and Observations

R = Randomization

ASP = Analytic Scoring Pattern Treatment on Experimental Group One

HSP = Holistic Scoring Pattern Treatment on Experimental Group Two.

Then the study covered all the public secondary schools in Obollo Afor Education Zone out of 5 other education zones of Enugu, Udi, Nkanu, Awgu and Nsukka of Enugu State. The choice of this zone is that many of the schools offer geography with substantial number of geography teachers. The population for this study consists of all the 47 Geography teachers and 337 senior secondary 3 (SS3) students in 46 senior secondary schools in Obollo Afor Education Zone.

In consideration of the fact that only secondary school Geography teachers were used for the study, a sample of 12 of the teachers were selected and used for test administration and scoring. The selection were done through a systematic sampling technique, where 17 schools were selected on 1 out of every 3 basis, starting from serial number 194 – 237 of Enugu State list of all public secondary schools (PPSMB, 2018). The geography teachers were randomly designed to experimental Group I and Experimental Group II. Only SS3 classes were used, having completed their scheme of work which covers all components of the essay test from SS1 to SS3. All the SS3 Geography students totaled 94 in number from the sampled schools were used for the exercise. However, the Geography teachers/raters were randomly assigned to Experimental

The researcher adopted a Geography Essay Test (GET) with scoring guide for the study.

The researcher adopted the 1999 West African Senior School Certificate Examination. Having all the psychometric properties of standardized tests, the researcher selected 2 and 3 essay questions each from the physical and human geography, respectively representing two broad divisions of geography. This however ensured its reliability, and validity.

The Geography teachers were trained on the 2 experimental treatment groups: analytic and holistic and were randomly assigned into 2 groups which were either experimental group I, or II that represented either analytic or holistic scoring pattern. The researcher trained the scorers on the two scoring patterns, after which they were used for test administration and scoring.

The researcher was satisfied with the scorers' performances and discussed how the research GET would be conducted. 93 students participated/responded in/to the GET, the scorers were issued with scripts, each with a required marking guide. 31 scripts (responses) selected for rating were duplicated into four such that either analytic or holistic raters had 62 scripts that are duplicated from the 31 scripts/responses. The 62 scripts were further divided into two sets each for the final assessments. The scorers were finally reduced to four of 2 each to Analytic or Holistic scorers, randomly selected through balloting without replacement techniques. The scorers were reduced

British International Journal of Education And Social Sciences
Vol.7, No.12; December- 2020
ISSN (3342 – 543X);
p –ISSN (4519 – 6511)
Impact factor: 8.11

in order to reduce minimally the individual subjectivity that may increase when many scorers are involved.Data collected were arranged and analyzed according to the research questions that guided the study. Mean and standard deviation were used to obtain the associate descriptive statistics, while Spearman's Rank Order Correlation Coefficient, was used to obtain the reliability estimates of within groups.

Results

The results of the study are presented in line with the research questions and hypotheses that guided the study.

Research Questions;

1. What is the influence of holistic scoring pattern on scorer reliability in Geography essay test?

Table 1: *Descriptive Statistics of two raters: Holistic vs Holistic*

Pattern	Mean	Std Deviation	N
Holistic	61.46	2.3483	31
Holistic	58.78	2.2013	31

Table 1 shows the statistical calculations regarding the condition in which all raters adopted a holistic pattern of scoring students' Geography essay test. That is to say, the two raters were asked to evaluate the same student essay test using holistic scoring guide. In line with the study regulations in order to meet up with the reliability and validity criteria, the same 31 essay tests were scored first by one rater and then by one other rater: in total by 2 holistic raters. **Table 1** indicates the descriptive statistics that include the means for both conditions (X=61.46 and M=58.78 respectively), standard deviations (SD=2.34 and SD=2.20 respectively), and the number of participants (N=31).

2. What is the influence of analytic scoring pattern on scorer reliability in Geography essay test?

Table 2: *Descriptive Statistic of two raters: Analytic vs Analytic*

	Mean	StdDev	N
Analyti	60.8711	2.324B	31
Analytic	61.3543	2.475	31

Table 2 shows the statistical calculations regarding the condition in which all raters adopted an analytic pattern in scoring of students' Geography essay test. That is to say, the two raters were divided into two and each was asked to elevate the same students essay based on analytic marking scheme. The first scorer was acquired to use a written analytic marking scheme containing 5 components (20 marks for each of the 5 questions) with 4 properties as a criterion whose sum of individual scores equals 100 points. In line with the study regulations in order to meet the reliability and validity criteria, the same 31 essay were also scored again by one other rater: I total by 2 analytic raters. Table 2 show descriptive statistics from the 31 essay scores. Table 2 shows the score means (M=60.87 and M=61.35 respectively) and the standard deviations (SD= 2.324 and SD=2.475 respectively) by two rating conditions: analytic versus analytic.

Hypotheses

Hypotheses 1: Holistic scoring pattern does not significantly influence scorer reliability in Geography essay test.

Table 4: *Correlation coefficients between two raters: Holistic vs. Holistic*

			Grade 1	Grade 2
Spearman's rho	Grade1-Holistic	Correlation Coefficient	1.000	0.763(**)
		Sig. (2-tailed)	.	0.000
		N	31	31
	Grade2-Holistic	Correlation Coefficient	0.763(**)	1.000
		Sig. (2-tailed)	0.000	.
		N	31	31

** Correlation is significant at the 0.05 level (2-tailed).

British International Journal of Education And Social Sciences
Vol.7, No.12; December- 2020
ISSN (3342 – 543X);
p –ISSN (4519 – 6511)
Impact factor: 8.11

Table 4, Provides the correlation coefficient of scores by all holistic raters in two rating conditions: holistic versus holistic. As indicated in Table 4, despite the fact that the score means (M= 61.46 and M= 58.78 respectively) are relatively close to each other, the statistical calculations in Table 4 show a high correlation coefficient (r= 0.763) between the essay scores given by two groups of holistic raters whose *p* value is also statistically significant (p=.000). In other words, the Spearman Correlation Coefficient Statistical Test reveals a highly positive correlation (r= 0.763) and inter-rater reliability between first scoring and second scoring of essays by two groups of raters who adopted a holistic style towards writing evaluation.

Hypotheses 2: Analytic scoring pattern does not significantly influence scorer reliability in Geography essay test

Table 5: *Correlation coefficients between two raters: Analytic vs Analytic*

			Grade 1	Grade 2
Spearman's rho	Grade1-Analytic	Correlation Coefficient	1.000	0.699(**)
		Sig. (2-tailed)	.	0.000
		N	31	31
	Grade2-Analytic	Correlation Coefficient	0.699(**)	1.000
		Sig. (2-tailed)	0.000	.
		N	31	31

** Correlation is significant at the 0.05 level (2-tailed).

In Table 5 the Spearman's correlation coefficient displays a moderate correlation of coefficient r= 0.699 with a statistically significant *p* value (p= 0.000). Literature regard a correlation between 0.70 and 1.00 as a high correlation, it can be observed that the obtained correlation (r= 0.699) is very close to the high correlation range.

Findings

Inter rater reliability of holistic vs. holistic scoring pattern on Geography essay test

Findings show that the correlation coefficient (r = 0.763) is highest between two pairs of Geography raters when they use a holistic rubric; that is to say, a highly positive inter-rater reliability is provided when two groups of raters score the same essays in terms of both the overall quality and their personal impression. Specifically, in a condition in which 31 Geography essay test were scored first by one holistic rater and subsequently again by one more holistic rater for reliability and validity purposes in accordance with exam regulations; a high correlation of 0 .763 emerged between these two rater pair. Since a correlation value between 0.70 to 0.89 has a high level of statistical significance according to Nworgu (2016), the holistic versus holistic essay rating condition is strongly correlated.

Inter rater reliability of analytic vs. analytic scoring pattern on Geography essay test

The second experimental condition of analytic vs. analytic seems to be corresponding to the viewpoint that if two groups of raters use a common marking scheme for essay scoring, their assessment results are normal to be in close agreement with another. Thus findings of the study indicated a positive correlation coefficient of r = 0.699, (very close to 0.70). Because of the criterion-referenced feature of analytic scoring pattern, it is considered to be superior, particularly for detailed scoring of an essay. Though not as high as holistic versus holistic condition, it draws attention to significant inter-rater reliability, In this context, the 31 student Geography essay test were rated by two groups of analytic raters (each consisting of 1 rater) who were given a detailed written marking scheme. Specifically, the 31 essays were first divided between 2 analytic raters so that each rater was assigned to score 31 essays in accordance with a specified component. Upon the completion of the first scoring, the same 31 essays were again given to 1 other analytic rater for a second scoring to help ensure reliability and validity standards. The

statistical calculations of the final average scores provided by the two groups of analytical raters, contrary to the common assumption, did not produce as high a correlation coefficient as the holistic versus holistic condition. Findings of this study revealed a moderate inter-rater reliability between analytic raters. Hence, this finding is similar to those made by Lee, Gentile and Kantor (2008), that six analytic scores were not only correlated among themselves but also correlated with the holistic scores, in a study on Analytic Scoring of TOEFL CBT Essays.

Conclusions

According to the statistical results of this study, out of the two conditions: (a) analytic vs. analytic (b) holistic vs. holistic. The holistic versus holistic condition produced surprisingly the highest level of correlation and accordingly inter-rater reliability. This finding was unanticipated, for the researcher given that literature is replete with hypothesis that the analytic versus analytic rating condition (which included raters who favoured detailed and comprehensive marking scheme for essay scoring) would always lead to more inter-rater reliability with grading by decreasing the range and variability among scores. Therefore, the results of this study put to question the common assumption that the adoption of analytic marking scheme in essay scoring highly increases the reliability or validity of writing evaluation. Hence, the statistical results of the holistic versus holistic condition are in support that holistic approaches may have enable raters to consider, in a consistent fashion, other qualities of writing than those specified in an analytic component.

Recommendations

1. Because of its practicality, affordability, and simplicity, holistic scoring pattern is recommended to schools. Even though, there is a lower inter-rater reliability of analytic condition, the noteworthy aspects of analytic marking scheme, like offering comprehensive information about the test taker's performance should not be shadowed by holistic rubric.
2. Given the unprecedented findings of this study, scoring of Geography essay should be more concerned in sharing the same rating scale based on a mutual understanding and interpretation.
3. Rather than being skilled in synthesis, raters may need to be adept in assessment through integration of both specified and unspecified written component
4. Training for a holistic scoring session should concentrate in modulating individual, idiosyncratic and subjective experience, so that raters can consistently perceive that whole.
5. In high-stakes situations, adopting holistic pattern, all papers and not just a sample, should be multi-scored, for three reasons. First, multiple holistic scores of student work will enable individual student performances to be independently rated by at least three markers to ensure that the mark is accurate and has been fairly assigned. Second, averaging ratings provides a better estimate of the true performance of the student because the error of measurement is reduced owing to replication. And third, multiple scoring generates marks carried to at least one decimal point to allow for more precise rank-ordering of students such as the Final Cumulative Grade Point Average (FCGPA), university entrance, or educational placement and promotion. Thus, multiple-scoring of all papers, which will compromise holistic scoring's cost-effectiveness, is necessary for both precise and fair decision making when the future of individual students is being considered.

References

Akintade, B.O. (2011). Considering the determinants of selecting geography as a discipline: The case of senior secondary school students in Ilorin, Nigeria. Ozea Journal of Social Sciences 3 (4): 131-138.

Brennan, R. (2001). An essay on the history and future of reliability from the perspective of replications.

Journal of Educational Measurement, Vol. 38, 295-317.

Brown GIL SLrving. S E, and Keegan PJ (2014) An introduction to Education assessment, measurement and evaluation, improving the quality of teacher based assessment (3rded). Auckland, ONZ. Dumore. Pub. ISSN 97819272120197

Çetin, Y. (2011).Reliability of raters for writing assessment: analytic - holistic, analytic - analytic, holistic – holistic. Q *Mustafa Kemal University Journal of Social Sciences Institute.Vol.8(16). Pp: 471-486*

Dawson. P (2015) Assessment Rubrics towards clearer and more replicable Design, Research and practice: Philip; Assessment and Evaluation in Higher Education doi: 10:1080/ 02602938:2015

Ebuoh, C.N. (2018): Effects of analytical & holistic hattern on scorer reliability in biology essay test. *Worldly Journal of Educational 8(1) 111-117.*

Institute of Australian Geographers (2020) Incorporated: Australian University. ABN: 97471418446 http://www.iag.org.au

Klein , S. P., Stecher, B. M., Shavelson, R. J., McCaffrey , D., Ormseth, T. and Bell , R. M. (2009). Analytic Versus Holistic Scoring of Science Performance Tasks. Journal of Applies Measurement in Education. Vol. 11(2).Pp: *121-137*

Lee, Y., Gentile, C. and Kantor, R. (2008).*Analytic Scoring of TOEFL® CBT Essays: Scores from Humans and E-rater*. ETS, Princeton, NJ

Madu, B. C. Ikeh, E. F. (2013).Effect of Scoring Patterns on Scorer Reliability in Economics Essay Tests. *Journal of Economics and Sustainable Development.Vol.4, (15).*

Metruk, R. (2018). Comparing holistic and analytic ways of scoring in the assessment of speaking skills. *The journal of teaching English for specific and academic purposes. Vol. 6, (1).Pp: 179-189*

Nworgu, BG (2006) Introduction to educational measurement and evaluation. Theory and practice(2nded). Nsukka. Hallman publishers

Nworgu, BG (2015). Educational measurement and evaluation. Theory and practice (2nded) Nsukka Nigeria. University Trust publishers

` Pidwirny, M. and Jones, S. (2016). Introduction to Physical Geography. University of British Columbia.Okanagan.physicalgeography.net.

Panadero.E and Johnson, A (2013) The use of scoring Rubrics for formation Assessment purposes revisited: A review of Research 9(40) Educational

Post Primary School Management Board (PPSMB) (2018).

Rupp, A.A. and Pant, H.A. (2007): Validity Theory in Salkind, Neil J. (ed): Encyclopedia of Measurement and Statistics. SAGE Publishing.

Saritha, K. (2016). Rubric for English Language Teaching Research. *Research Journal of English Language and Literature. Vol. 4* (2), *725-731*.

Saxton, E., Belanger, S. and Becker, W. (2012) The Critical Thinking Analytic Rubric (CTAR): Investigating intra-rater and inter-rater reliability of a scoring mechanism for critical thinking performance assessments. *Assessing Writing. Vol. 17(4), 251-270.*

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Xi, X and P. Mollaun (2006).Investigating the Utility of Analytic Scoring for the TOEFL Academic Speaking Test (TAST).TOEFL iBT Research

Report.http://s3.amazonaws.com/academia.edu.documents/30193670/rr-06

Yaqub, H., Tabassum, R. and Farooq, M. (2016).Intra-rater reliability of holistic and rubric-based assessment of essay writing in Pakistan. *Science International.(Lahore), Vol.28 (4). Pp: 669-680,2016*