Journal of Economics and Sustainable Development
ISSN 2222-1700 (Paper) ISSN 2222-2855 (Online)
Vol.4, No.15, 2013

Effect of Scoring Patterns on Scorer Reliability in Economics Essay Tests

Madu, B. C. Ph.D*, Ikeh, Elochukwu Francis
Department of Science Education, University of Nigeria, Nsukka Nigeria.
*Email of the corresponding author: bcmadu4owa@gmail.com orbarnabas.madu@unn.edu.ng.

Abstract
The study investigated the effect of scoring patterns on scorer reliability in Economics essay test. In this study, one research question was posed and one hypothesis was also tested. The sample of the study comprised of all the 20 Economics teachers and 120 Senior Secondary II (SS2) Economics students from the public secondary schools in Aguata Education Zone in Anambra State, Nigeria. Economics Essay Test (EET) was used for data collection which was administered to 120 SS2 Economics students to generate the scripts used for the study. The instrument (EET) was trial tested using 20 SSII Economics students who share the same characteristics with the study subjects in order to determine the reliability of the instrument. The reliability coefficient of 0.89 was obtained using scorer reliability formula. The data obtained for the study were analyzed using Kendall's coefficient of concordance (w) in answering research question, while chi-square test of significance of Kendall's coefficient was used in testing the null hypothesis. The findings of the study revealed that scoring an item across board was more effective in scoring Economics essay test. Based on the finding it was recommendation that scoring an item across board should be adopted for improving scorer reliability of Economics essay test in both internal and external examination; it is also necessary that the scoring pattern recommended should be incorporated in the curriculum of teacher training institution since the use of this scoring pattern in schools is not popular
Key words: Measurement, Assessment,. Scoring patterns, Examinations, Essay questions, Economics.

1.Introduction

Measurement of learning outcome in Education is carried out through assessment. Assessment is the process of gathering information about students' abilities or behavior for the purpose of making decisions about the students (Elliot, Kratochmill, Cook and Travers, 2000). Different assessment formats are utilized by classroom teacher depending on the objective of the measurement. These assessment formats for determining the students' understanding of key course topics include multiple-choice questions, true-false, fill-in-the blanks, shorts answer, problem solving exercises and essay questions. Most of the alternatives to multiple-choice and true-false questions are described in literature as constructed response or essay test questions, meaning that they require students to create their own answers rather than select the correct one from a list of pre written alternatives. Essay test is one of the assessment tools utilized by classroom teachers especially when the teacher wants the students to originate, organize, express, and integrate ideals in a given problem (Agwagah, 1997). Essay test is described as one or more essay questions administered to a group of students under standard conditions for the primary purpose of collecting evaluation data. Essay questions are usually categorized into two types namely extended response questions and restricted response questions (Mehrens and Lehmann 1978). Extended response questions (ERQ) allow students freedom to determine the content and to organize the format of their answer; the students decide which facts are pertinent and how to organize, synthesis, and evaluate them. Such questions are most appropriate when the objective is to test student's writing skills, including conceptualization, organization, analysis, synthesis, and evaluation, giving the student minimum or ample choice regarding the topic.
Restricted response questions (RRQ) are questions that limit both the content and the form that the students answer may take. Restricted response questions are the appropriate form required for testing the content. For this study, the researchers adopted the restricted response questions for writing economics essay test. Failure to establish adequate and effective limits for the student response to the essay question allows students to set their boundaries for their response, meaning that students might provide responses that are outside, the intended task. If students' failure to answer within the intended limits of the essay question can be ascribed to poor or ineffective wording of the task, the teacher is left with unavailable and invalid information about the students' achievement of the intended learning outcome and has little or no basis for grading the student responses. The essay tests have the following objectives (Cashin, 1987).

- They can be used to test learning outcomes not measurable by other means.
- They can test thought processes, the students' ability to select, organize, and evaluate ideas etc; and their ability to apply, integrate, think critically and solve problems.
- Require that students use own writing skills; the students must select the words, compose the sentences and paragraph, organize the sequence of exposition, decide upon correct grammar and spelling.

- Pose a more realistic task than multiple-choice and other "objective" items. Most of the life's questions and problems do not come in a multiple-choice format, and almost every occupation requires people to communicate in sentences and paragraphs, if not in writing.
- Cannot be answered correctly by simply recognizing the correct answer; it is not possible to guess.
- Can be constructed relatively quickly.

These objectives indicate that educators choose essay questions over other forms of assessment because essay items challenge students to create a response rather than to simply select a response. Some educators use them because essays have the potential to reveal students' abilities to reason, create, analyze, synthesize, and evaluate. Hence, essay tests can be used to assess higher-order or critical thinking skills, evaluate students' thinking and reasoning skill; provide authentic experience, test thought process (Reiner, Bothell, Sudweeks, and Wood , 2002), and it takes relatively little time to construct and minimizes guessing. However, despite these advantages of essay test, there is much subjectivity in its scoring due to the fact that students organize their responses to questions in different ways and hence allocating marks for essay-type or free-response questions can be very unreliable (Baird, Greatorex, & Bell, 2004), and therefore it is said to have low scorer reliability (Onunkwo, 2002).Scorer reliability or inter-rater reliability, inter-rater agreement or concordance is the degree of agreement among the raters.(Gwet, 2010). Scorer reliability is the degree of correspondence or agreement among the scores given to students by different examiners (Abonyi, 2011). It gives a score of how much homogeneity or consensus there is in the ratings given by judges or raters. It is useful in refining the tools given to individuals judges and for determining if a particular scale is appropriate for measuring a particular variable given by a single score.

It is common to observe variations among the scores received by different individuals. Some of this variation is due to actual differences in the characteristic being measured; however, there may be other factors that contribute to the observed variation. These other factors are sources of error that prevent an accurate assessment of the object of measure. Reliability is an expression of the proportion of the variation among scores that are due to object of measure. As variation due to error goes to zero, the reliability of an assessment goes to 1. Factors that may serve as sources of error in an essay test include;

- Variations in the students writing proficiency.
- Variations in the content knowledge ie domain expertise, of those evaluating the essay test, and
- The consistency with which the essay tests are evaluated.

This third factor depends in large part on the selection of method/pattern by which essay tests are scored. The effect of the selection of a scoring pattern on the assessment scorer reliability is of primary concern of this study. The first two factors, subjects writing proficiency and raters domain expertise, were assumed to contribute little to the variation of essay test scores. Scoring patterns according to Ebuoh and Okafor (2011) are various methods that are employed by scorers to obtain the quantitative performance of students. Some of the scoring patterns as reported by Ebuoh and Okafor (2011) are as follows:

- Scoring all the items at the same time.
- Scoring an item across board.
- Ranking of all scripts before scoring all items.
- Division of task scoring into section.

Scoring all the items at same time presupposes that the teacher scores all the questions for one student before picking another script. If in an Economics test for instance, the students were requested to answer five questions, the teacher has to mark all five questions for a student before entering into another script. This does not facilitate fast marking and yet, it is the popular marking style that most teachers adopt (Onunkwo, 2002).

Scoring an item across the board is the pattern in which the teacher scores one question for all the students completely before the teacher enters into another question. If it was question number one (1) for example that the teacher starts with, the teacher has to finish marking this question for every student who took the examination before picking up another question. This makes scoring faster since it enables the teacher concentrate on a question and its making scheme at a time. By the time the teacher has scored about twenty (20) students' responses to a particular question, the teacher must have become very familiar with all the points in the marking scheme and their associated marks regarding the question.

Ranking all script before scoring all items is another pattern of scoring essay test in which the answers are not divided into points to which marks are awarded. There is just one standard answer to each question. The teacher (scorer) reads each student's response to a question, compares it with the standard answer, and awards a grade or mark he considered appropriate based on his pre-set criteria. The scorer may use numerical grades (eg, 80%), letter grade (eg, A, B, C) or comments (eg, above average, superior quality, poor work, excellent work, below average, etc).

Division of task of scoring into session pattern is also one of the scoring patterns employed especially where large papers are to be scored. In this pattern, one scorer specializes in scoring a section (a part or number) of the

test. The scorer scores his section and passes on the script to the next scorer. Evidences abound in support of this method, for example, Lovegrove (1984) advised that in an examination in which crucial decisions may be taken, two or more scorers should be allowed to score a section of the script independently.

However, an attempt to compare these patterns and determine the one that will yield higher scorer reliability of scoring Economics essay test is therefore the concern of this study since little efforts have been made to compare these scoring patterns and determine the one that enhances higher scorer reliability when employed in scoring essay tests in Economics.

The reliability of essay test depends on how well it is been scored (Okpala, 2003). This shows that there is a problem of scoring essay test even when the same marking scheme is used, and this can also affect students' academic achievement positively or negatively, in the sense that student responses can be over scored or under scored without reflecting the actual performance of the students being assessed. Though they may be factor(s) that may affect the scorer reliability of essay test, but Mehrens and Lehman (1978) observed that a carefully planned, constructed and administered essay test can be ruined by improper scoring pattern and standards. This calls for the need for proper scoring pattern in scoring essay test which Economics is no exception.

1.1 Purpose of the Study

The main purpose of this study is to investigate the effect of different scoring patterns on scorer reliability in Economics essay tests.

1.2 Research Question

The research question posed to guide this study is:

- What is the effect of scoring patterns on scorer reliability in Economics essay test?

1.3 Research Hypothesis

The null hypothesis (H_0) formulated and tested at 0.05 level of significant for this study is:

- There is no significant difference in the correlation coefficients of scoring patterns of scorers who scored Economics essay tests using different scoring patterns.

2.Methodology

2.1 Design.

Under this design 20 scorers in each condition (Scoring all the items at the same time, scoring an item across board, division of task scoring into section, ranking of the scripts before scoring) scored the full set of 120 writing responses of senior secondary two students of economics. This resulted in a completely crossed design in which every scorer scored every paper. This resulted in a total of 20 raters at each class equally distributed among the four scoring conditions. Each prompt was scored for both content and conventions. The final score was obtained by combining the content and convention score so that the content score was given twice the weight as the convention score. The study design is shown in figure 1

Scoring conditions				
Scorers	**Scoring all the items at same time.**	**Scoring an item across board**	**Ranking of all scripts before scoring all items.**	**Division of task scoring in sections**
Scorers	**1, 2, 3, 4, 5, … 20**	**1, 2, 3, 4, 5, … 20**	**1, 2, 3, 4, 5, … 20**	**1, 2, 3, 4, 5, … 20**
1				
2				
3				
4				
.				
.				
.				
120				

Figure 1. Study design for scoring.

We further examine the data from this study to better understand the trends in the findings as utilized by kreiman, 2007. In our analysis we were interested in the differences between the ratings of the 20 scorers in each condition and the specifically, we calculated mean differences (bias), standard deviation of differences and the total root mean square errors, using the following

$$\text{Bias} = \frac{1}{N_p}\sum_{i=1}^{N_p}\left(Di\right)$$

Where Di = rater score for paper i – Researcher score for paper i

N_p = total number of papers.

$$SD_{diff} = \sqrt{\frac{1}{N}\sum_{i=1}^{N_p}(Di - Bias)^2}$$

Root mean square error RMSE = $\sqrt{Bias^2 + SD_{diff}^2}$

2.2 Study Area

The study was conducted in Aguata Education Zone of Anambra State of Nigeria. Aguata Education Zone is made up of three (3) Local Government Areas namely; Aguata, Orumba South, and Orumba North, with 47 public secondary schools. The sample size for this study was all the 20 Economics teachers and 120 SSII Economics students. All the 20 Economics teachers in the public secondary schools in Aguata Education Zone were used for this study because the population is manageable. On the hand, two secondary schools were purposively selected which made up of 120 SSII Economics students.

The instrument used for data collection was Economics Essay Test (EET) developed by the researchers. The Economics Essay test (EET) was based on the following Economics topics: Demand and Supply, concept of money, Agriculture, Distributive trade, and Production drawn from SSI&2 syllabuses. The test was developed by preparing table of specification based on the six levels of cognitive domain of Bloom's taxonomy of education. The test contains five items (questions) with sub questions in each item, and each item carries an equal mark of (20) totaling (100) in the whole test. The researchers also developed scoring guide for the scoring of the developed Economics essay test.

The instrument was face-validated by four experts in Economics. These experts were required to examine the items with respect to:

- Whether the items constructed correspond as indicated in the table of specification.
- The structure and clarity of the items.
- Whether the answers to the questions correspond or tally with those provided in the scoring guide.
- Whether the language used in constructing the questions is suitable for SSII Economics students.

The comments and recommendations of these experts served as a guide to the modification of items in the EET. For content validity, the use of a well constructed and validated table of specification was used in constructing the EET items. The instrument was trial tested using fifty (50) SSII Economics students who share the same characteristics with the study subjects to obtain the internal consistency of the instrument and it was found to be 0.89.

For controlling extraneous variables in this study, the following measures were adopted:

2.3 Training of teachers (scorers)

There was training programme of all the teachers (scorers) that were involved in the scoring. Scorers are recruited based on requirements developed. Depending on the assessment, specific educational and experience requirements are met (ie degree, experience as a classroom teacher). Scorers were trained using comprehensive training materials developed by the researchers. During this period, the validated instrument and scoring guide were discussed. Teachers pay more attention to the processes and procedures utilized for constructed response item scoring. Although specific constructed response procedures vary slightly based on the teachers conventions and requirements, certain components are universally addressed including pattern development, range finding, scorer selection procedures, scorer training and qualification, and scorer monitoring

Pattern development for constructed response items is carried out at the beginning of a programme. Depending upon the content and type of pattern either a single rule may be developed and applied to all items, or item-specific rules may be developed and in small cases, a holistic rule may be developed for a content area with item specific rules for each item. Once constructed response items are developed and field-tested, range finding is carried out.

Range finding is the process used to determine how to apply the rule to students' papers and therefore determine the standards that are used to score constructed essay response items. The process may also define both a range of response types or performance levels within a score point on the rule and the threshold between score points. This implies that the researchers determine where one score point ends and another begins. The range finding process defines the papers that are characteristic of the various score points represented by the rule. The researchers look at a pool of responses which cover the range of score points for a particular item and through scoring and discussion come to an agreement score on each response. The researchers made notes from discussion and used these notes for interpreting the responses to use to train scorers. The researchers assured that the scorers are scoring based on the standards set when responses with consensus scores are used as anchor and training papers. For instance, some sample of constructed responses were scored twice, providing the basis for a number of statistics related to inter-rater reliability, for instance, perfect agreement, perfect plus adjacent

agreement, spearman correlation, kappa statistics etc).
In addition, papers previously scored by researchers were distributed to scorers and form the basis for validity indices that are similar to the inter-rater reliability statistics. Validity papers were chosen specifically because of certain features that can test, for instance, whether scorers are consistently applying the consensus-based logic to borderline paper. Scorers were monitored in a variety of ways. For instance scorers were also monitored using back-reading, whereby a research leader rescored papers from a certain scorer or scorers that are performing at marginal level of reliability. Through this process, the research leader provided specific feedback or additional training in real time. There was also trial scoring of the dummies by the scorers during the training exercise.
For data collection, the instrument (EET) was administered to 120 SSII Economics students by the researchers to generate the scripts that were distributed to Economics teachers (scorers) who scored the scripts using different scoring patterns after the scorers have been randomly assigned to the different scoring patterns. After scoring of the scripts by the scorers, the researchers personally collected them for recording and analysis.
Data collected was analyzed using Kendall's coefficient of concordance (w) in answering the research question while chi-square test of significance of Kendall's coefficient was used in testing the null hypothesis. Trends in the findings were also analyzed by determining the bias, standard deviation of difference and root mean square error.

3 .Results

Research Question 1**:** What is the effect of scoring patterns on scorer reliability in Economics essay test?

Table 1**:** Summary of Kendall's coefficient of concordance (w) of the four scoring patterns

Scoring Patterns	No of Scorers	Kendall's Coefficient (W)
Scoring all the items at the same time.	20	.23
Scoring an item across board.	20	.78
Ranking of all scripts before scoring all items.	20	.65
Division of task scoring into section	20	.71s

From table 1 above, it was revealed that scoring an item across board had a positive relationship with a high correlation coefficient of 0.78 while division of task also had a positive relationship with correlation coefficient of 0.71. Ranking before scoring recorded a correlation coefficient of 0.65 while scorin all the items at the same time pattern had a low correlation of 0.23.these coefficient indices indicate the level of agreement among the raters in each pattern.

Hypotheses

There is no significant difference in the correlation coefficient of the scoring patterns of scorers who scored Economics essay tests using the four different scoring patterns.

Table 2: Test of significance of Kendall's coefficient of the scores awarded by scorers in the four different scoring patterns

Scoring patterns	N	W	df	X^2	P
Scoring all the items at the same time.	20	.23	19	30.14	0.15
Scoring an item across board	20	.78	19	30.14	0.00
Ranking of all scripts before scoring all items.	20	.65	19	30, 14	0.00
Division of task scoring into section	20	.71	19	30.14	0.00

Table 2 above shows that the correlation coefficients for scorers who scored Economics essay test using scoring an item across board (SAIAB), dividing the task of scoring into sessions pattern (DSISP), and ranking all script before scoring all items (RASBSAI) were significant. This is due to the fact that the exact probability value of P is less than .05.. In the same vein, the exact probability value of P is greater than .05. This shows that the use of SAIAB, DSISP, and RASBSAI had a significance difference on scorer reliability in scoring Economics essay test. Because of these findings, any apparent pattern may be due simply to chance which is further analyzed in table 3.

Table 3 summarizes the means of the bias, standard deviation of differences and RMSE values across all scorers in each condition.

S/N	**Content Score**		
	Bias	SDDIF	RMSE
1	-0.17	0.70	0.74
2	-0.25	0.67	0.71
3	-0.23	0.66	0.70
4	-0.23	0.68	0.72

The scoring an item across the board condition had the lower overall mean scorer bias compared with the other three conditions across all scores. However, the scoring of all the items at the same time condition has the highest mean deviations of differences and overall means RMSE. This provides some additional detail about the results of the study. The distribution was artificial and represented much higher mean than the national distribution of scores in the full sample of papers.

4.Discussion

Researchers evaluated the results of the study in terms of inter-rater agreement among the scorers. Their results were inconclusive in that same statistically significant differences between the scoring conditions were found. Whereas there was a suggestion of a pattern in the data, this study found no consistent statistically significant differences in reliability between distributed scoring and traditional local scoring. The findings of the study show that there is effect of scoring patterns on scorer reliability in Economics essay test. The study reveals that scorers who scored Economics essay test using scoring an item across board pattern of scoring recorded a high scorer reliability coefficient followed by division of task of scoring into session, ranking all scripts before scoring all items, and scoring all items at a time. The essence of reliability is to ensure free error measurement, therefore the relative superiority of scoring an item across board over other patterns of scoring in enhancing scorer reliability in Economics essay test could be attributed to the fact that, by the time the teacher has scored about 20 students' responses to a question; he must have become very familiar with associated marks regarding the question. This is why Onunkwo (2002) argued that scoring an item across board promotes scorer reliability. This implies that Economics teachers are more consistent in using scoring across board in scoring than any other pattern. The finding was further strengthened using chi square test of significance of Kendall's coefficient which reveals that the relationship was significant for scorers who scored Economics essay test using scoring an item across board (SAIAB), dividing the task of scoring into sessions pattern (DSISP), and ranking all script before scoring all items (RASBSAI). The results of this study disagrees with the findings of Ebuoh and Okafor (2011) who reported that ranking an item before scoring was more effective scoring pattern in scoring Biology essay test. However this study has proven that scoring across board is more reliable in scoring Economics essay test compare to other scoring patterns. In many studies comparison of student responding to essay response items by paper-and-pencil have been confounded by potential rater effects in rating hand written essays. When essays are handwritten, scorers may give writers the benefit of the doubt if spelling or punctuation, for instance, is not clear. This benefit of the doubt may vary from rater to rater. Just as the style, tidiness, size or any other characteristic of student handwriting can inappropriately influence the scorers assigned by raters, raters must be constantly reminded to evaluate student responses based on the author paper and scoring guides rather than the easy or difficulty of reading responses. That is, with reference to table 3, it is likely that regression to the mean effects contributed to the negative bias. A second reason was that the monitoring process (eg validity checks, back-reading etc) that typically occurs in operational scoring was suspended for the study. This was done because it was thought that the monitoring process would contaminate the comparison across conditions. The patterns of scorer error are very similar across the four conditions of study. Breland, Lee and Muraki (2005) speculated that the most plausible expectation for these findings seems to be that typed essays are likely to be perceived by scorers as final drafts, and therefore expectations are slightly higher and errors in grammar or spelling are judged mere harshly than they are in handwritten essays. In many studies, comparisons of students responding to constructed response items by paper-and –pencil have been confounded by potential rater effects in rating hand written essays

5.Conclusion

On the basis of the findings in this study, it was found that scoring an item across board was found more effective in enhancing scorer reliability of Economics essay test and there is no significant difference in the correlation coefficient of scoring pattern of scorers who scored Economics essay tests using scoring all the items at the same time pattern. Hence,since scoring an item across board were found more effective in improving scorer reliability of Economics essay test, it is recommended for its use in scoring Economics essay test in both internal and external examination and should be incorporated in the curriculum of teacher training institutions.

References

Abonyi, S. O. (2011). *Instrumentation in Behavioural Research*: A Practical Approach. Enugu: Timex Publisher.

Agwagah, U.N.V. (1997). Types of test-essay and objectives. In S.A. Ezeudu, U.N.V. Agwagah & C.N. Agbaegbu (eds). *Educational Measurement and Evaluation for Colleges and Universities.* Onitsha: Cape publishers International Limited.

Baird, J.A., Greatorex, J., & Bell, J. P. (2004). What makes marking reliable? Experiments with UK Examinations. *Assessment in Education*, 11(3), 331-348.

Breland, H., Lee, Y.E. & Muraki, E.(2005). Comparability of TOEFL CBT essay prompts: respons mode analysis. *Educational and Psychological Measorement,65(4),577-595.*

Cashin, E.N.(1987).Improving essay test. IDEA PAPER NO.17.Center for Faculty of Evaluation and Development, Kansas State University.

Ebuoh, C.N. & Okafor, G.A. (2011). Effects of scoring by session, ranking and conventional patterns on scorer reliability in Biology essay tests. *International Journal of Educational Research.*11 (1) 242-252.

Elliot, S. N., Kratochill, T. R., Cook, J. L., and Travers, J. F. (2000). *Educational psychology: Effective teaching, effective learning* (3rd ed). Boston: McGraw Hill.

Gronlund, N. E. (1976). *Measurement and evaluation in teaching* (3rd ed.). New York: Macmillan publishing Co., Inc.

Krieman, C. (2007). Investigating the effects of training and rater variables on reliability measures: A comparison of standup local scoring, online distributed scoring, and online local scoring. Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.

Lovegrove, M. N. (1984). Evaluating the result of learning. In B. O. Ukeje (ed). *Foundations of Education.* Benin City: Ethiope Publishing Corporation.

Mehrens, W.A. & Lehmann, I.J. (1978). *Measurement and evaluation in education & psychology.* (2nd ed). New York: Holt Rinehart and Winsten Inc.

Nworgu, B.G. (2006). *Introduction to Educational Measurement and evaluation: theory and practice* (2nd ed.). Nsukka: Hallman Publisher.

Okpala, D. (2003). Standardization, test administration and scoring. In B. G. Nworgu (ed), *Educational measurement and evaluation: theory and practice* (136-142). Nsukka: University Trust Publisher.

Onunkwo, G. I. N. (2002). *Fundamentals of educational measurement and evaluation.* Owerri: Cape Publishers Int'l Ltd.

The West African Examination Council (2011). Vetting sheet on Economics.

Gwet, K.L (2010*). Handbook of Inter-rater Reliability* (2nd edition)

http://www.agreestat.com/book_exercepts.html retrieved 20th september,2012.