



## Relative Effectiveness of Marking Experience on Reliability of Holistic and Analytic Scoring of Essay Test in Economics

Onyeagba, Mercy C. N. & Adinna, Patricia I.

Department of Educational foundations, University of Nigeria, Nsukka

Corresponding Author: [mercy.onyeagba@unn.edu.ng](mailto:mercy.onyeagba@unn.edu.ng)

### Abstract

The study focused on the relative effectiveness of Marking Experience of examiners on Reliability of Holistic and Analytic Scoring of Essay Test in Economics Anambra State, Nigeria. Correlational research design was used in the study. Two research questions and two null hypotheses were formulated and tested at .05 level of significance. The population of 255 economics teachers in public secondary schools in Anambra State was used in the study. A sample of 128 economics teachers representing 50% of the population was used. It consists of 64 economics teachers for holistic scoring methods and 64 economics teachers for analytical scoring methods. The two scoring methods were explained and randomly given to the two groups, each consisting of 64 scorers (one group to one method), to score 28 examination scripts. The statistics used was Pearson Product Moment Correlation Coefficient statistics. The findings showed that as marking experience increases that SEM decreases in analytical and holistic scoring methods but the difference is statistically significant in holistic scoring at .05 level of significance. Based on the above findings the researcher suggested that since training matches experience in marking reliability, that intensive training in the use of holistic scoring methods should be given to less experienced teachers, to put them on the same pedestal with the more experienced teacher for the realization of the hope of a better marking reliability with the use of scoring methods

**Key words:** Marking experience, Reliability, Holistic scoring, and Analytic scoring.

### Introduction

Experience is the knowledge gained through repeated trials (Tom, 2017). Experience is a particular instance of personally encountering or undergoing through something. Experience is also an event or series of events participated in or lived through. Akpan (2017) defined experience as a process of gaining knowledge or skill. Therefore, experience is the professional growth that takes place in an individual as a result of continued stay, training and retraining on the job and other related processes. In the context of this study, experience is regarded as the number of years a marker has marked. Some markers are highly experienced while others are averagely experienced. The less experienced markers are likely to mark less consistently than highly experienced ones. The study by Barkaoui, (2010) indicated that novice raters were more lenient than experienced raters in giving more importance to argumentation, while the experienced raters were more severe regarding grammatical accurateness. Similarly, there are number of studies that showed that inexperienced markers were more severe and/or less accurate than experienced markers. For example, Baker, Ayres, O'Neil, Chli, Sawyer, Sylvester and Carroll (2008) found some evidence that experience of marking a state-level high-stakes examination improved the

reliability with which Australian teachers marked key stage 3 English tests. However, this pattern has not been found in all studies and, where differences occurred, they could often be negated by training (Meadows & Billington, 2011).

Unfortunately, human judges have many opportunities to introduce error in scoring of essays (Linn & Gronlund, 2006). Barkaoui, further revealed that rubric types had more effect on the scoring processes than rater experience. The variations in the rubrics affected inexperienced raters more than experienced raters. Researchers have also diverse view about experience including marking experience. Their argument centered on whether marking experience is more reliable with holistic or analytic scoring (Barkaoui, 2010; Cumming, 2003; Pinot de Moira, 2003; Murphy, 1988).

Holistic scoring method is a method of scoring students' composition in which one or more evaluators read the paper without marking anything, and then rate the paper as a whole, assigning single scores based on total impression of composition as a whole text or discourse. The holistic scoring method is based on the theory that a whole piece of writing is greater than sum of its parts (Bacha, 2011). In other



words, holistic rubrics assess students' work as a whole.

Analytic scoring is a method of evaluating students' performance that involves assigning a separate score for each aspect of a task (Murphy, 1988; Cumming, 2003). It is usually used to assess students' worth on individual basis. It identifies and assesses aspects of a finished product (Pinot de Moira, 2003).

Despite concerns about fairness and consistency in scoring, there is little empirical research that systematically compares the reliability of the two approaches to scoring essay test. Even though there are significant number of studies connected with each rubric type in particular holistic and analytic, there is an unfortunate scarcity of experimental research that aims to compare the effectiveness of these rubrics in the same context including inter rater reliability (Brookhart, 2005). Marking criteria are often not investigated to see whether they work and raters are not often trained to give consistent marks. Some of the assessment used in schools may have high face validity-they may look excellent to the uniformed but they may be marred by inappropriate marking schemes and rating inconsistencies. This may produce invalid results which are unfair to students and teachers alike. The importance of experience in marking has been highlighted by many researchers but much work has not been done on the relationship existing between marking experience and scoring methods. Therefore, there is need for consideration of the association between marking experience and the scoring methods which will be of great importance to the production of valid results which are fair to students and teachers. Against this background therefore, the main thrust of this study is to determine the relationship existing between marking experience and scoring methods (analytic and holistic). Based on this, the following questions were posed to guide the study:

1. What is the relationship between marking experience and standard error of measurement in the scores awarded by the group that used holistic scoring method?
2. What is the relationship between marking experience and standard error of measurement in the scores awarded by the group that used analytical scoring method?

Also based on the research questions, these hypotheses were formulated

**HO<sub>1</sub>:** There is no significant relationship between marking experience and standard error of measurement in the scores awarded by the group that used holistic scoring method.

**HO<sub>2</sub>:** There is no significant relationship between marking experience and standard error of measurement in the scores awarded by the group that used analytical scoring method.

### Research Method

The study employed the correlational research design. The research adopted this design based on the recommendation of Maduabum (2004) that, correlational research design can be used in research when it aims at determining the relationship between variables. This research design was adopted because it typically allows the researcher to ascertain the association between marking experience and the scoring methods.

This study was conducted in Anambra State, Nigeria. The State is located in the South East geopolitical zone of Nigeria. The people of Anambra State are mainly traders, public servants, entrepreneurs and artisans. Anambra State comprises six Education Zones namely Aguata, Awka, Nnewi, Ogidi, Onitsha and Otuocha education zones. The State has a total of 257 State government owned secondary schools as at 2018 (Post Primary Schools Service Commission (PPSSC), Awka, 2018). The State is also a business-oriented area, as a result, majority of teachers get involve in different businesses while holding unto their teaching jobs. Consequently, this portends great challenge to teachers, especially those in the secondary schools and perhaps in scoring of examination papers.

The population of the study was 255 economics teachers in public secondary schools in Anambra State (Post Primary School Service Commission (PPSSC) Awka, 2018)). Their marking experiences ranged between 1-30 years. Thus, every Economics teacher in Anambra State public secondary school is a member of the population. Simple random sampling technique was used to select a sample of 128 Economics teachers, representing 50% of the population. It consisted of 64 economics teachers for analytical scoring method and 64 economics teachers for holistic scoring method.

The instrument for the study was the two-marking scheme that is, holistic and analytic, for two essay test items in economics (essay questions for 2017/2018 WASSCE). The researcher adopted the



instrument from WASSCE Coordination Center for marking (2017/2018) Senior Secondary School Certificate Examination. The marking scheme was screened and discussed by the 40 scorers, for corrections and necessary adjustments during the coordination meeting, prior to the marking session, thus, the collective conclusions ensured the validity of the marking schemes/guides. The reliability was carried out through the test-retest method. The study employed 12 teachers to score the scripts of Community Secondary School, Okpuno, Awka, according to marking scheme using the analytic and holistic scoring methods. The scripts were retrieved and scored after an interval of two weeks. The same 12 markers were given the same marking scheme to score the same scripts using the analytic and holistic scoring methods. On analysis of first and second sets of results, using Pearson Product Moment Correlation Coefficient technique, the reliability coefficients of .82 and .82 for analytic and holistic scoring methods were obtained respectively. A .82 overall reliability indicated that the study is highly reliable.

The data for the study were obtained through the two groups of 64 scorers each, who scored the same 28 scripts. The marks awarded by the 128 scorers constituted the data for the study. The marks/scores were used to compare the holistic and analytical scoring methods. Each marker was given a parcel containing 28 clean photocopies of the students' answer scripts in the essay test, mark sheet and the question paper. Each parcel had one of the methods of scoring and specified years of experience labeled on it. These parcels were distributed to each examiner according to their years of experience. The two groups were separated, they all sat differently. The researcher explained the two methods and the purpose of the research. The marking guide was then discussed, item by item. Corrections that were accepted by all were effected in the guide and the corrected version was used for the scoring. The scoring exercise was coordinated by the researcher and two briefed assistants.

To answer research question 1 and 2 which addressed the association between marker's experience and the marks awarded by the two groups of markers; the researcher correlated the scores awarded by individual markers with their years of marking for WASSCE in the two groups, using the Pearson Product Moment Correlation Coefficient statistic. The linear relationship between SEM and

marking experience in the two groups were also computed.

**Table 1: Pearson Product Moment Correlation between the Marking Experience and Standard Error of Measurement in the Scores Awarded by the Group that Used Holistic Scoring Method**

N	R
64	-.452

The result of Pearson Product-Moment Correlation Coefficient from the Table 1 showed the relationship between marking experience and standard error of measurement in the scores awarded by the group that used holistic scoring method with:  $r = -.452$  and  $n = 64$ . The Table 1 revealed a correlation coefficient of  $-.452$  which indicated that there is a moderate but negative relationship existing between marking experience and standard error of measurement in the scores awarded by the group that used holistic scoring method. This means that as marking experience increases that SEM decreases also and vice versa.

**Table 2: Pearson Product Moment Correlation between the Marking Experience and Standard Error of Measurement in the Scores Awarded by the Group that used Analytical Scoring Method.**

N	R
20	-.125

The result of Pearson Product-Moment Correlation Coefficient from the Table 2 above showed the relationship between marking experience and standard error of measurement in the scores awarded by the group that used analytical scoring method with:  $r = -.125$  and  $n = 20$ . Table 2 revealed a negative correlation coefficient of  $-.125$  which indicated that there is a negative but weak relationship existing between marking experience and standard error of measurement in the scores awarded by the group that used analytical scoring method. The negative correlation of marking experience and standard error of measurement in analytical scoring implies that as years of marking experience increases, the SEM occurrence decreases and vice versa.

**Table 3: Pearson Product Moment Correlation Coefficient between the Marking Experience and Standard Error of Measurement in the Scores**



**Awarded by the Group that used Holistic Scoring Method**

N	R	Sig. (2-tailed)
64	-.452	.035

The result of Pearson Product-Moment Correlation Coefficient from the Table 3 above showed the significant relationship between the marking experience and standard error of measurement in the scores awarded by the group that used holistic scoring method with:  $r = -.452$ ,  $n = 64$  and  $p\text{-value} = .035$ . Since  $p\text{-value} (.035)$  is less than  $0.05$ , the study rejects the null hypothesis and accepts the alternative hypothesis that there is a significant relationship between the marking experience and average mark change in the scores awarded by the group that used holistic scoring method.

**Table 4: Pearson Product Moment Correlation Coefficient between the Marking Experience and Standard Error of Measurement in the Scores Awarded by the Group that Used Analytical Scoring Method**

N	R	Sig. (2-tailed)
20	-.125	.019

The result of Pearson Product Moment Correlation Coefficient from the Table 4 above showed the significant relationship between the marking experience and standard error of measurement in the scores awarded by analytical scorers with:  $r = -.125$ ,  $n = 20$  and  $p\text{-value} = .019$ . Since  $p\text{-value} (.019)$  is less than  $.05$ , the study accepts the null hypothesis. Therefore, there is no significant relationship between the marking experience and average mark change in the scores awarded by the group that used analytical scoring method.

**Discussion of Findings**

**Marking Experience (MEX) and Standard Error of Measurement (SEM) in the Scores Awarded by the Group that Used the Holistic Scoring Methods (HSM)**

Findings on marking experience and standard error of measurement in the scores awarded by the group that used the holistic scoring methods revealed a moderate negative correlation coefficient. This indicated that there is a negative relationship existing between marking experience and standard error of

measurement in the scores awarded by the group that used holistic scoring method. The moderate negative correlation of marking experience and standard error of measurement in holistic scoring implies that when there are more years in marking experience, that SEM reduces and vice versa. The study also showed that there is a significant relationship between the marking experience and average mark change in the scores awarded by the group that used the holistic scoring method. The negative correlation could be as a result of scorers' inexperience in holistic scoring methods. Murphy (1988) and Shohamy, Gordon & Kraemer (2005) are of this same view with the study. Similarly, the level of SEM in the holistic scoring method in this study is also the result of its higher level of acquaintance to marker's inexperience. Pinot de Moira (2003) is of the opinion that markers' inexperienced are more prone to random errors in holistic marking.

**Marking Experience (MEX) and Standard Error of Measurement (SEM) in the Scores Awarded by the Group that Used the Analytical Scoring Methods (ASM)**

Findings on marking experience and standard error of measurement in the scores awarded by the group that used the analytical scoring method revealed a low negative correlation coefficient. This indicated that there is a negative relationship existing between the marking experience and standard error of measurement in the scores awarded by the group that used analytical scoring method. The weak negative correlation of marking experience and standard error of measurement in analytical scoring implies that as years of marking experience increases, that the SEM occurrences reduces and vice versa but the relationship is a weak one. The study also showed that there is no significant relationship between the marking experience and average mark change in the scores awarded by the group that used the analytical marking method.

Conclusively, the marks awarded by the analytical and holistic scorers as shown in this study were negatively affected by markers' inexperience. This point is supported by studies which show that inexperienced markers are less consistent, that is more prone to random errors in marking than experienced markers (Murphy, 1988; Pinot de Moira, 2003).



### Recommendation

Based on the findings of the study, the following recommendations were made:

1. Examination bodies should ensure adequate training for inexperienced markers since marking experience tends to reduce standard error of measurement (SEM) in analytic and holistic scoring.
2. The Ministry of Education needs to organize, in service training as well as seminars for teachers so as to abreast their knowledge in scoring methods.
3. Examination bodies should employ more experienced markers than inexperienced markers during marking coordination.
4. More focus should be placed on experience as it is more reliable and SEM tends to occur less.

### References

- Akpan, H.F. (2017). *Administration of School Community relations and principals' effectiveness in cross River State, Nigeria*. (Ph.D thesis). University of Calabar.
- Bacha, N. (2011). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System*, 29(3), 371-383.
- Baker, E., Ayres, P., Sawyer, W., Sylvester, R. M. & Carroll, B. (2008). KS3 English test marker study in Australia: Final report to the National Assessment Agency of England. Sherman Oaks, CA: University of Southern California O'Neil, H. F.
- Barkaoui, K. (2010b). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, 7(1), 54-74.
- Brookhart, S. M. (2005). The quality of local district assessments used in Nebraska's school-based teacher-led assessment and reporting system (STARS). *Educational Measurement: Issues and Practice*, 24, 14-21.
- Cumming, A. (2003). Expertise in evaluating second language compositions. *Language Testing*, 7(1), 31-5
- Linn, R. L. & Gronlund, N. E. (2006). *Measurement and assessment in teaching*. (7th edition) Englewood Cliffs, NJ: Prentice-Hall.
- Maduabum, M. A. (2004). *Fundamentals of Educational Research*. Owerri: Versatile Printers
- Meadows, M. & Billington, L. (2011). A review of the literature on marking reliability, report to National Assessment Agency.
- Murphy, R. J. (1988). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology*, 52(1), 58-63
- Pinot de Moira, A. (2003). Examiners' background and the effect on marking reliability. *AQA Research Report, RC218*.
- Shohamy, E., Gordon, C. & Kramer, R. (2005). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, 76(10), 27-33.
- Tom, T.S (2017) performance of tests of teaching proficiency: national development and validation. *America educational journal*, 8, 105-117