SUB-THEME

COLLABORATION FOR HARMONIZATION OF EDUCATIONAL ASSESSMENT STANDARDS

ASSESSING THE CONSISTENCY OF EXAMINERS' SCORING IN ENGLISH LANGUAGE AND MATHEMATICS OF NABTEB CERTIFICATE EXAMINATION

DR. MOHAMMED A. MOHAMMED
E-MAIL: mammf_2005@yahoo.com

MR. PIUS S. OSAIGBOVO
E-MAIL: osaigbovopius70@gmail.com

DR. PHILOMENA E. IRO-AGHEDO
E-MAIL: philiroaghedo@gmail.com

DR. CATHERINE I. OMEONU
E-MAIL: kateomeonu@gmail.com

NATIONAL BUSINESS AND TECHNICAL EXAMINATIONS BOARD (NABTEB), BENIN CITY, NIGERIA

A PAPER PRESENTED AT THE 41ST AEAA ANNUAL CONFERENCE HOLDING IN ADDIS ABABA, ETHIOPIA FROM 25TH TO 29TH AUGUST, 2025

Introduction

- The credibility and integrity of any examination body largely depend on the consistency and accuracy of the scores awarded to candidates.
- The consistency of examiners in marking candidates' scripts in any examination is fundamental to maintaining the validity and reliability of assessment outcomes.
- In high-stakes assessments such as those conducted by examination bodies like NABTEB, it is crucial that examiners apply uniform standards in evaluating students' scripts.
- The importance placed on marking of candidates' scripts becomes a priority to all assessi Boards since the results from these examinations support national planning and policy-making.

Introduction Contd.

- To maintain fairness, NABTEB strive to ensure that candidates receive accurate scores that reflect their true performance through inter-rater reliability.
- Inter-rater reliability refers to the consistency between different raters when evaluating candidate's work or assigning scores.
- Inter-rater reliability indicates the extent to which independent raters obtain the same results using the same rating criteria to rate an examinee's performance (Popham, 2011).

Introduction Contd.

There are a number of statistics that have been used to measure inter-rater reliability, these includes;

- Percentage agreement
- Cohen's kappa (for two raters)
- Pearson r and the Spearman Rho
- Contingency coefficient
- Fleiss kappa (adaptation of Cohen's kappa for 3 or more raters)

Introduction Contd.

- Cohen's Kappa (κ) is a widely used statistic for assessing inter-rater reliability.
- Kappa values range from –1 to +1, where 0 indicates no agreement and 1 signifies perfect agreement between raters.
- Kappa values below 0 are possible, Cohen notes they are unlikely in practice and suggest worse than chance agreement, indicating substantial disagreement among raters (Marston, 2010).
- Cohen's Kappa is a standardized statistic, ensuring consistent interpretation across various studies.

Introduction Contd.

Various factors can impact the consistency and accuracy of scoring in subjective assessments.

- Sandler (2009) noted that markers often exhibit varying tendencies: some are notably lenient, others are more stringent and some display inconsistencies in their assessments.

- Weiner (2007) opined that factors which influence inter-rater reliability include: subject to be observed, raters, atmosphere in measurement time and the instrument.

Other researchers reported factors such as

- assessors' fatigue (Sehar & Mahmood, 2020),
- individual biases, varying levels of experience (Madu & Ikeh, 2018).
- subjective interpretations of assessment criteria (Gladkoff, Han & Nenadic 2023).

Introduction Contd.

- These above mentioned studies collectively underscore the various nature of scoring discrepancies between examiners.

- The implication is that when high stakes examinations are marked by a panel of examiners, the examiners should be coordinated in a way that ensures no candidate gains an unfair advantage or disadvantage based on the examiner who assesses their work.

Statement of the Problem

- The accuracy and fairness of examination results depend largely on the consistency and reliability of examiners' scoring particularly in core subjects like English Language and Mathematics.
- To maintain fairness, NABTEB strives to ensure that candidates receive accurate scores that reflect their true performance.
- The quality assurance measures put in place include vetting of examiners' scoring to correct for deviations from the accepted scores in the marking scheme and checking scores summation.
- Despite these measures, the scoring of candidates' scripts continues to pose a potential source of error that can affect the accuracy and reliability of test scores.
- This study therefore assessed the consistency of examiners scoring in English Language and Mathematics in the National Business Certificate (NBC)/National Technical Certificate (NTC) examinations using inter-rater reliability.

Research Questions

- What is the level of inter-rater agreement between the English Language Team Leaders and Assistant Examiners in 2024 In-School NBC/NTC marking exercise? 1
- What is the level of inter-rater agreement between the Mathematics Team Leaders and Assistant Examiners in 2024 In-School NBC/NTC marking exercise? 2
- What are the factors contributing to deviations in scores in English Language and Mathematics? 3

Methodology

Research Design

- Mixed research design, incorporating both ex-post facto and survey research approaches.

Population

- All 480 examiners involved in the marking of English Language and Mathematics for the 2024 In-School NBC/NTC marking exercise in Nigeria.

Sampling Technique and Sample

- Purposive sampling technique was used to select 232 English Language and 216 Mathematics examiners.

Methodology Contd.

Instrument

- English Language and Mathematics Vetting Sheets.
- Semi-structured Questionnaire on Marking and Scoring of Scripts Variation.

Data Analysis

- Inter-rater reliability was established using Cohen's Kappa Statistics.
- Data from the semi- structured Questionnaire was analyzed thematically.

Methodology Contd.

Interpretation of Cohen's Kappa statistics (k) value

- ≤ 0 = No Agreement
- 0.10 - 0.20 = Slight Agreement
- 0.21 - 0.40 = Fair Agreement
- 0.41 - 0.60 = Moderate Agreement
- 0.61 - 0.80 = Substantial Agreement
- 0.81 - 0.99 = Almost Perfect Agreement
- 1.00 = Perfect (McHugh, 2012)

Results

1 What is the level of Inter-rater Agreement between the English Language Team Leaders and Assistant Examiners in 2024 In-School NBC/NTC marking exercise?

Table 1: Symmetric Measures showing Kappa values of Inter-rater Agreement between the English Language Team Leaders and Assistant Examiners

S/N	Marking Venue	No. of Vetted Scripts	Kappa Value (κ)	Asymp. Std. Error	Approx. Tb	Approx. Sig.	Inter-Rater Agreement
1	Ado-Ekiti	285	0.650	0.029	64.143	0.000	Substantial
2	Asaba	127	0.723	0.041	45.127	0.000	Substantial
3	Awka	307	0.688	0.027	68.659	0.000	Substantial
4	Benin	609	0.576	0.021	78.921	0.000	Moderate
5	Enugu	665	0.552	0.020	78.407	0.000	Moderate
6	Ibadan	298	0.516	0.030	42.796	0.000	Moderate
7	Ilorin	447	0.662	0.023	86.205	0.000	Substantial
8	Katsina	296	0.809	0.023	87.862	0.000	Almost Perfect
9	Lagos	281	0.854	0.021	87.164	0.000	Almost Perfect

Table 1: Contd.

S/N	Marking Venue	No. of Vetted Scripts	Kappa Value (κ)	Asymp. Std. Error	Approx. Tb	Approx. Sig.	Inter-Rater Agreement
10	Makurdi	244	0.529	0.033	45.209	0.000	Moderate
11	Minna	359	0.623	0.026	73.707	0.000	Substantial
12	Osogbo	302	0.656	0.028	69.335	0.000	Substantial
13	Owerri	276	0.562	0.030	59.209	0.000	Moderate
14	Sokoto	161	0.852	0.029	56.938	0.000	Almost Perfect
15	Uyo	203	0.673	0.033	66.095	0.000	Substantial
16	Yola	279	0.562	0.030	57.947	0.000	Moderate
Overall		5139	0.639	0.007	291.203	0.000	Substantial

2 What is the level of Inter-rater Agreement between the Mathematics Team Leaders and Assistant Examiners in 2024 In-School NBC/NTC marking exercise?

Table 2: Symmetric Measures showing Kappa values of Inter-rater Agreement between the Mathematics Team Leaders and Assistant Examiners

S/N	Marking Venue	No. of Vetted Scripts	Kappa Value (κ)	Asymp. Std. Error	Approx. Tb	Approx. Sig.	Inter-Rater Agreement
1	Ado-Ekiti	246	0.900	0.019	82.729	0.000	Almost Perfect
2	Asaba	410	0.965	0.009	135.859	0.000	Almost Perfect
3	Awka	310	0.652	0.027	84.011	0.000	Substantial
4	Benin	643	0.846	0.014	143.904	0.000	Almost Perfect
5	Enugu	681	0.819	0.015	116.216	0.000	Almost Perfect
6	Ibadan	401	0.918	0.014	104.336	0.000	Almost Perfect
7	Ilorin	408	0.905	0.015	127.693	0.000	Almost Perfect
8	Katsina	203	0.980	0.010	91.233	0.000	Almost Perfect
9	Lagos	191	0.818	0.028	80.007	0.000	Almost Perfect

Table 2: Contd.

S/N	Marking Venue	No. of Vetted Scripts	Kappa Value (κ)	Asymp. Std. Error	Approx. Tb	Approx. Sig.	Inter-Rater Agreement
10	Makurdi	421	0.804	0.020	96.617	0.000	Almost Perfect
11	Minna	455	0.937	0.011	143.447	0.000	Almost Perfect
12	Osogbo	436	0.860	0.017	134.921	0.000	Almost Perfect
13	Owerri	452	0.723	0.021	98.391	0.000	Substantial
14	Sokoto	364	0.885	0.017	125.156	0.000	Almost Perfect
15	Uyo	230	0.715	0.030	71.840	0.000	Substantial
16	Yola	293	0.996	0.004	93.808	0.000	Almost Perfect
Overall		6144	0.856	0.005	470.172	0.000	Almost Perfect

3 What are the factors contributing to deviations in scores in English Language and Mathematics?

- An English Language examiner has this to say on subjectivity in marking:

  Some of the examiners are subjective in scoring. They are carried away by candidates' handwriting and touching story in comprehension and thereby ignoring the grammatical errors.

- In a similar vein, an examiner identifies fatigue as one of the causes of deviation in scoring and responded thus:

  Examiners mark continuously to meet up with the deadline as such they get tired and are no longer consistent.

- Another remark from a Mathematics examiner that inexperience is one the factors for deviation as expressed below:

  Inexperience examiners always ignore symbols and units when marking which is not supposed to attract full mark.

Discussion of Findings

- The inter-rater reliability between the Team Leaders and Assistant Examiners in English Language was Substantial Agreement and statistically significant.

  Finding supports Kayapınar (2014); Gladkoff, Han & Nenadic (2023) who reported that there is always variation in evaluating Writing Skill and the reliability of scoring even if evaluators are experienced linguists.

- Also, the inter-rater reliability between the Team Leaders and Assistant Examiners in Mathematics was Almost Perfect Agreement and statistically significant.

  This is consistent with Pantzare (2015) who find high inter-rater reliability of teachers' ratings of national tests in Mathematics irrespective of the reliability of the measure used.

Discussion of Findings Contd.

- Factors identified responsible for scores deviation are examiners' inexperience, fatigue, non-adherence to marking scheme, lack of adequate training of examiners during coordination exercise and lack of commitment on the part of examiners.
- This result supports Meadow and Billington (2013) who noted that compared to experienced markers, inexperienced markers tend to mark more severely and employ different rating strategies.
- It is also consistent with Sehar & Mahmood (2020) who stated that raters' fatigue could affect the reliability of markers.
- Uwadiae and Oke (2018) who maintained that the most important aspect of good and accurate marking is the marker's familiarization with the marking scheme.

Conclusion

- The study revealed that the inter-rater reliability between the Team Leaders and Assistant Examiners in English Language was Substantial Agreement and statistically significant.
- Also, the inter-rater reliability between the Team Leaders and Assistant Examiners in Mathematics was Almost Perfect Agreement and statistically significant.
- The study identified subjectivity in scoring, Inappropriate Marks Allocation, Time constraints, Fatigue over time, Emotional Stability, Examiners' inexperience, Non-adherence to marking scheme, Lack of adequate training of examiners during coordination exercise and Lack of commitment on the part of examiners as partly responsible for variation in scores.

Recommendations

- NABTEB should recruit competent and well-trained examiners to participate in the making exercises.
- Subjectivity should be reduced through the development of comprehensive marking scheme.
- Time constraints and examiners' fatigue should be mitigated by managing the number of assigned scripts effectively.
- Strict adherence to marking schemes must be enforced and appropriate sanctions for non-compliance.

THANK YOU FOR LISTENING