International Journal of Humanities, Education, and Social Sciences

e-ISSN : 3026-0892
p-ISSN : 3026-1422

Index: Harvard, Boston, Sydney, Dimensions, Lens, Scilit, Semantic, Google, etc

https://doi.org/10.58578/IJHESS.v4i2.9109

Application of Intra-Class Correlation to Multi-Rater Agreement on Students' Grading

Adetunji A. A., Alo N. G., Anibijuwon A. A.
Federal Polytechnic, Ile-Oluji, Nigeria
adeadetunji@fedpolel.edu.ng

Article Info:

Submitted:	Revised:	Accepted:	Published:
Feb 9, 2026	Mar 25, 2026	Apr 9, 2026	Apr 14, 2026

Abstract

Fair and credible student assessment depends on consistent grading practices, particularly when multiple raters evaluate academic performance using a shared rubric. This study aimed to examine the consistency of students' grading by applying the intraclass correlation approach to multi-rater assessments of academic performance. The reliability of ratings was evaluated using the Intraclass Correlation Coefficient (ICC) for continuous rating scores. Data were obtained from records of students' assessments of project and seminar presentations. The findings revealed varying levels of agreement among raters, indicating that grading practices were generally inconsistent, although some degree of agreement was observed across different assessed variables. These results highlight the importance of applying inter-rater reliability statistics in educational assessment to identify inconsistencies in scoring practices. The study concludes that improving grading fairness and objectivity requires regular rater calibration and the consistent use of standardized rubrics. This study contributes to assessment practice by emphasizing the value of reliability analysis in strengthening the quality and credibility of academic evaluation.

IJHESS Journal is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

Keywords: Inter-Rater Reliability; Intraclass Correlation Coefficient; Grading Consistency; Academic Assessment; Standardized Rubrics

INTRODUCTION

Multi-rater agreement is the degree of consensus or consistency among multiple raters, evaluators or observers who are assessing, evaluating or rating the same set of objects or subjects (Ibe & Onuoha, 2020). It also measures the extent to which different raters agree with each other when evaluating the same phenomenon (Nwankwo & Ahmed, 2020). Such an agreement is essential and often used in different fields, including Education and Medicine. In academic environments, especially at the tertiary level, the accuracy and reliability of grading systems are paramount for maintaining fairness, credibility, and academic integrity (Jones *et al.,* 2023). Grades influence students' academic progression, scholarship opportunities, and future careers. Therefore, inconsistencies in grading due to human subjectivity and evaluator bias are of significant concern in educational measurement and evaluation (Adebowale & Ogunyemi, 2021).

Using multiple raters improves the reliability and validity of the grading system. However, it is essential to assess the agreement among raters to ensure that grades are consistent and reliable. Differences in interpretation or judgment among multiple grades can lead to biased results (Wang *et al.,* 2020). The issue of grading consistency becomes more critical in courses that involve subjective assessment components such as essays, projects, and oral presentations (Musa & Ibrahim, 2024). In such contexts, the presence of multiple raters or assessors introduces variability that may distort the true performance of students. Thus, assessing inter-rater reliability becomes necessary to ensure that the grading process is equitable and valid (Chukwuemeka & Odili, 2022).

This study investigates the level of agreement among multiple raters that grade students' academic work in project and seminar presentations. The study uses the intra-class correlation coefficient to assess the reliability of grades across different evaluators. Without strong agreement among raters, students' evaluations may reflect individual rater tendencies more than actual student achievement (Nwankwo & Ahmed, 2020). The intra-class correlation coefficient (ICC) is widely utilized and employed in determining inter-rater reliability (Boateng & Amponsah, 2020). The ICC is suitable for continuous data and

provides information on the degree of correlation and agreement among raters (Banjo & Etim, 2022). In recent years, the use of the ICC in academic settings has increased, particularly in evaluating the fairness of grading in departments where multiple instructors assess students. Studies have shown that even with clearly defined rubrics, discrepancies in grading still exist, necessitating the need for quantitative assessment of rater agreement (Musa & Ibrahim, 2024). In Nigerian polytechnic settings, where practical and subjective assessments are common, the application of rater agreement statistics has been underutilized. This calls for more context-specific researches to evaluate how consistent academic assessments are in such institutions (Okoro & Eze, 2022). Ensuring that grades awarded are consistent across different evaluators is essential to support institutional credibility and uphold the validity of student outcomes. Discrepancies in scores from different instructors can undermine the evaluation process. Tools like the ICC can help identify inconsistencies and suggest improvements to the grading process (Yusuf, 2023).

The development and adoption of a standardized marking scheme does not automatically translate to consistent grading. Human interpretation, mood, bias, and other cognitive factors still influence rater decisions. It is therefore vital to empirically assess the extent to which multiple graders agree when evaluating the same work (Oladipo & Ayeni, 2020). The study of inter-rater reliability using statistical techniques not only identifies inconsistencies but also contributes to professional development by highlighting the need for calibration among educators (Fatoba & Ajayi, 2020). It enables departments to improve grading systems, introduce moderation practices, and enhance the overall quality of education (Adeola & Ajayi, 2025).

Assessment is a key pillar of academic evaluation, and grading is widely used to evaluate students' achievement. Inconsistent grading among evaluators is a persistent challenge in tertiary institutions. The problem of inconsistent grading among raters can lead to unfair assessment of students' performance, which can have serious consequences. While subjectivity is sometimes inevitable in assessment, the lack of empirical evaluation of rater agreement means grading may be arbitrary in some situations. This study investigates how well multiple raters agree when grading students' performance, utilizing the ICC. This leads to the following research questions:

i. What is the level of agreement among raters on student scores using ICC?

ii. How consistent are raters when scores are graded using ICC?

iii. Are there notable differences in rater agreement across different assessment types?

The study contributes to the growing body of knowledge on educational assessment and reliability within Nigerian polytechnics. The results can serve as a foundation for improving assessment procedures, including standardization of rubrics and training of assessor. Additionally, this research encourages accountability and fairness in student evaluation, which is vital for academic integrity and students' motivation.

Literature Review

Conceptual Review

Multi-Rater Agreement

Multi-rater agreement is fundamental in educational settings, where assessments such as essays and presentations, often involve evaluation by more than one rater. The consistency in these evaluations reflects the fairness and credibility of the grading process (Okoro & Eze, 2022). Variations in judgment can occur due to differences in experience, interpretation of marking schemes, or even unconscious bias when students are graded by different lecturers. High inter-rater agreement ensures students' performance is judged based on established standards rather than subjective interpretation. Conversely, low agreement points to a need for clearer rubrics, rater training, or better alignment of grading criteria (Adeola & Ajayi, 2025). Analysis involving multi-rater agreement helps in implementing moderation processes, where grades are reviewed and harmonized before final release (Adebowale & Ogunyemi, 2021).

Intra-class Correlation Coefficient (ICC)

In educational research, the ICC is used to evaluate the degree of similarity or agreement among multiple quantitative measurements or ratings when different raters score the same set of students' performances, such as in project or report evaluations (Nwankwo & Ahmed, 2020). It quantifies how much of the total variation in scores is due to differences between subjects versus differences between raters. A high ICC value (closer to 1) indicates excellent agreement among raters, while a low value (closer to 0) suggests poor agreement (Kwarteng, 2023).

The ICC can accommodate different rating designs, such as one-way random, two-way random or two-way mixed models, depending on the structure of the data and study design (Yusuf, 2023). The one-way random effects model, ICC (1, 1), is used to assess the

reliability of ratings or measurements made by multiple raters. The model assumes that the raters are randomly selected from a larger population of potential raters. For the two-way random effects model, ICC (2, 1), the assumption is that both the raters and the subjects/items being rated are randomly selected from larger populations while the two-way fixed effects model, ICC (3, 1), assumes that the raters are fixed, meaning they are the only raters of interest and the results are not intended to generalize to other raters.

In student assessments, the ICC helps to identify if grading differences are due to true performance variations or inconsistencies among raters. The value of ICC would be high if all raters consistently rate students similarly (Oladipo & Ayeni, 2020). Also, the choice of the ICC model and its interpretation must align with the nature of the study. For instance, the two-way random model is often appropriate when both raters and subjects are randomly selected, which can be useful in generalizing findings across populations (Adeola & Ajayi, 2025).

Grading Reliability in Educational Assessment

Reliable grading ensures that students' academic outcomes are reflective of their actual performance and not influenced by inconsistencies in judgment or measurement error (Adeola & Ajayi, 2025). Factors that influence grading reliability include the clarity of assessment rubrics, the training and experience of evaluators, the complexity of the subject matter, and the mode of assessment. Inconsistent use of grading criteria or unclear rubrics often leads to wide discrepancies in the scores assigned by different instructors (Chukwuemeka & Odili, 2022). When raters are trained on how to apply rubrics, the level of inter-rater agreement tends to improve significantly (Okoro & Eze, 2022). When students perceive the grading process as arbitrary or unfair, it can lead to disengagement, loss of motivation, and academic dissatisfaction. Thus, standardization and statistical evaluation of grading practices must be prioritized by institutions to promote transparency and fairness in student assessment (Adebayo & Eze, 2023).

Theoretical Review

Classical Test Theory (CTT)

This is one of the oldest and most widely used theories in educational measurement. The theory posits that any observed score (X) is made up of two components: (i) a true score (T), and (ii) an error score (E), expressed as $X = T + E$. This model is foundational in evaluating test reliability, including the reliability of grading systems (Adeola & Ajayi, 2025).

In this study, the CTT provides a basis for examining how much variability in students' scores is due to actual differences in performance versus inconsistencies introduced by raters. The smaller the error component, the more reliable the assessment is considered to be (Adebayo & Eze, 2023). While the CTT has been instrumental in the development of standardized testing, it has limitations in that it assumes all items and raters contribute equally to the error. In real-world assessments, especially those involving human judgment, different raters may vary significantly in their severity or leniency (Nwankwo & Ahmed, 2020). Nevertheless, the CTT lays the foundation for reliability indices such as Cronbach's Alpha, which parallels the use of ICC in measuring consistency across raters. It is also useful in explaining why discrepancies may exist in scores, even when the same criteria are applied (Chukwuemeka & Odili, 2022).

In grading systems where multiple evaluators assess the same work, the CTT encourages attention to the sources of error and promotes methods to control them. This includes double-marking, rater training, and post-assessment moderation; all strategies aimed at reducing error variance and increasing grading reliability (Okoro & Eze, 2022).

Generalizability Theory (G-Theory)

This theory extends the ideas of the CTT by considering multiple sources of measurement error. Unlike the CTT, the G-theory decomposes error into specific facets such as raters, items, and occasions. This makes it especially useful in multi-rater assessment contexts (Yusuf, 2023). The theory supports the application of statistical tools like the ICC by providing a framework for identifying how much variance is attributable to raters versus students. For instance, in a G-theory study, variance components are estimated to determine how each factor contributes to total score variability (Kwarteng, 2023).

The G-theory is particularly relevant in educational settings involving performance-based assessments, such as oral presentations, project reports, or portfolios. These assessment forms typically involve multiple raters, analyzing rater-based error critically (Adebowale & Ogunyemi, 2021). Its strength lies in its ability to guide decision-making and help educators determine how to improve reliability, whether by increasing the number of raters, refining rubrics, or using repeated assessments. This proactive use of reliability estimation is valuable for continuous improvement in assessment practices (Chukwuemeka & Odili, 2022).

Item Response Theory (IRT)

The IRT is a modern approach to educational measurement with a focus on the relationship between a test-taker's latent ability and their probability of giving certain responses. Though traditionally used in standardized testing, the IRT also offers insights into rater behaviour and scoring patterns (Musa & Ibrahim, 2024). In the IRT, the difficulty of test items and the ability of students are modelled separately, which helps isolate inconsistencies in scoring. In the context of rater agreement, extensions of the IRT, such as rater-effect models, allow the detection of bias, severity, and discrimination levels of individual raters (Okoro & Eze, 2022).

An advantage of the IRT is its ability to evaluate whether different raters are applying the same performance standards. For example, two raters may agree on which students pass or fail, but differ in how they interpret borderline performances. The IRT models can help reveal such hidden discrepancies (Adeola & Ajayi, 2025). Although the IRT requires complex statistical modeling and large datasets, it offers an advanced understanding of measurement fairness. It can be particularly beneficial in departmental evaluations where rater effects need to be minimized for high-stakes decisions such as graduation or scholarship eligibility (Nwankwo & Ahmed, 2020).

Empirical Review

Usage of ICC in Education

Several empirical studies have used the ICC to assess inter-rater reliability in educational settings, especially where multiple instructors or examiners are involved. A study by Oladimeji and Yusuf (2023) evaluated the consistency of 5 lecturers scoring student laboratory reports in a Nigerian polytechnic and reported an ICC value of 0.84, indicating excellent agreement. The researchers concluded that standardized rubrics and rater training contributed to the high reliability. In another study by Molefe and Van Rensburg (2021), the ICC was used to analyze the consistency among multiple assessors of final-year engineering projects. An ICC value of 0.69 was obtained, indicating moderate agreement.

In another study involving nursing students' clinical evaluations, Adebowale and Ogunyemi (2021) used the ICC to examine differences between clinical instructors. Their findings indicated that assessors with more experience showed better alignment in scores,

while novice assessors contributed to variability. The study recommended mentorship for new assessors to improve reliability. In a United Kingdom study by Wang *et al.* (2022), the inter-rater reliability was applied in essay scoring among school teachers. An ICC value of 0.72 was reported. The researchers noted that alignment increased significantly when assessors underwent collaborative marking sessions and used electronic rubrics.

In Nigeria, Adebayo and Ediale (2024) evaluated the inter-rater reliability in oral defence scoring among lecturers in a polytechnic in Ekiti State using the ICC. The study revealed inconsistencies among faculty members with an ICC value of 0.62, prompting a call for departmental moderation policies to ensure score harmonization.

Despite the popularity of ICC, its correct application depends on the design and selection of appropriate models (e.g., one-way random vs. two-way random). Studies have shown that misunderstanding the statistical assumptions behind ICC can lead to misinterpretation of results (Nwankwo & Ahmed, 2020).

Rater Reliability in Nigerian Institutions

Rater reliability is increasingly becoming a focus in Nigerian tertiary institutions due to growing concerns about fairness and transparency in academic evaluations. In a study by Okoro and Eze (2022), rater agreement was assessed in a College of Education in Enugu State across three departments. The study found wide variation in scores, with ICC values ranging from 0.58 to 0.81. This indicated inconsistencies in grading practices across departments and led to the implementation of centralized marking strategies. Similarly, Akinwunmi and Dada (2021) investigated grading discrepancies in practical examinations in a Nigerian university. Their research revealed that inconsistencies often stemmed from the absence of clear rubrics and poor inter-rater communication. The study recommended integrating digital rubrics in grading systems to standardize criteria.

A recent study by Omole and Adebiyi (2023) in a polytechnic in Ogun State analyzed final-year project grading and found that raters with different academic backgrounds (e.g., Engineering versus Science) tended to apply grading scales differently. This study highlights the role of discipline-specific perceptions in grading and calls for department-wide rater alignment workshops. Raters' biases in continuous assessment were analyzed by Fatoba and Ajayi (2020), using both the ICC and the Kappa statistics. The findings revealed a lack of training as a major factor contributing to unreliable grading. Their work informed the institution's decision to mandate rater calibration before high-stakes assessments.

Collectively, these studies, among many others, underscore the relevance of evaluating rater reliability in Nigerian higher education and validate the need for tools such as the ICC in identifying and correcting grading inconsistencies.

Impact of Rater Training on Scoring Consistency

Several empirical studies have highlighted the importance of rater training in improving inter-rater agreement. A study by Ede and Oduwole (2022) conducted in three Nigerian polytechnics assessed grading consistency before and after a structured rater calibration session. The results showed a significant increase in the ICC values from 0.61 to 0.85, confirming the effectiveness of training in harmonizing rater judgments. In another study conducted at Obafemi Awolowo University, Nigeria, Adebayo and Eze (2023) found that assessors who had undergone training produced significantly more consistent scores when grading English compositions. Cohen's Kappa increased from 0.48 to 0.73 after two training sessions focused on using rubrics and handling borderline cases. Also, Ajibola and Fatimah (2021) investigated the long-term effects of periodic rater workshops on inter-rater reliability. Their study revealed that consistency improved over time, and lecturers who attended at least three training sessions had significantly higher ICC scores than those who did not. The study recommends institutionalizing rater training at the start of each academic session.

A study in Ghana by Boateng and Amponsah (2020) found that training not only improved scoring reliability but also reduced marking time, as raters gained confidence and clarity in applying assessment criteria. This benefit can support the adoption of training programs in resource-constrained institutions. In contrast, Okeke and Olumide (2021) reported mixed outcomes, where training improved agreement in essay writing but had little effect in practical evaluations. This suggests that rater training should be tailored to the type of assessment and should involve discipline-specific examples to be most effective.

Beyond individual skills, training sessions also promote shared understanding among assessors. Overall, the evidence confirms that rater training is a key intervention for improving grading consistency. Institutions committed to fairness in assessment should prioritize continuous professional development for their teaching staff.

Technology-Assisted Assessment and Reliability Monitoring

In recent years, the integration of technology in educational assessment has created new avenues for improving grading reliability. Online rubrics, grading software, and learning

management systems (LMS) now allow for real-time monitoring of rater behaviour and scoring patterns. A study by Bello and Uche (2023) in Lagos State Polytechnic revealed that automated rubrics embedded in LMS platforms improved inter-rater consistency among instructors by over 20%.

Another empirical work by Oseni and Lawal (2024) at the University of Ilorin examined rater consistency when using electronic marking sheets in large undergraduate classes. The ICC values improved significantly compared to paper-based assessments. The researchers concluded that digital tools helped standardize input, reduce calculation errors, and prompt raters to follow scoring schemes more closely. However, in a critical review, Yusuf (2023) cautioned that unless digital platforms are designed with clear rubrics and user training, they may amplify inconsistencies rather than eliminate them. Moreover, internet issues and software familiarity can affect the smooth use of grading tools.

Despite these challenges, technology is rapidly reshaping how inter-rater agreement is assessed and improved. Studies agree that combining human oversight with automated tools yields better outcomes than either approach alone.

METHODS

Research Design

This study adopts a descriptive quantitative research design using secondary data. The data comprises multi-rater grading of all students in the Department of Statistics, Federal Polytechnic, Ile-Oluji, Nigeria, during seminar and project presentations of Year 2 (ND II) and Year 4 (HND II) for two academic sessions. At least two raters graded each student at all times. The Intra-class Correlation Coefficient (ICC) (for continuous grades) is used for inferential analysis. The assessed variables are: Introduction, Presentation, Physical Appearance, Practical Application, and Contribution to Knowledge.

Statistical Tools

Intra-class Correlation Coefficient (ICC) assesses the degree of agreement between raters on continuous scores. In this study, the two-way random effects model with absolute agreement is assumed. The Mean Square Between (MSB), the Mean Square Within (MSW) and the Mean Square Between Raters (MSR) are obtained using the following formulas:

$$MSB = \frac{\sum_{i=1}^{k}(n_i \times (X_i - \bar{X})^2)}{(K-1)}$$

$$MSW = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - X_i)^2}{N-K}$$

$$MSR = \frac{(\sum_{j=1}^{r}(r_j \times (X_j - \bar{X})^2)}{(r-1)}$$

where; n_i =number of observations in group *i;*

X_i = mean of group *I* $\bar{X}$ = grand mean

y_{ij} = individual observation k = number of groups

r_j = number of ratings per rater X_j = mean rating per rater

r = number of raters N = Total number of observations

To calculate ICC (2,1) using the following formula;

$$ICC(2,1) = \frac{[MSB - MSW]}{[MSB + (k-1) \times MSW + (r-1) \times MSR]}$$

The ICC value ranges from 0 (no agreement) to 1 (perfect agreement) and is interpreted as:

- < 0.40: Poor agreement among raters
- 0.40-0.59: Fair agreement
- 0.60-0.74: Good agreement
- 0.75-1.00: Excellent agreement

These interpretations are easily influenced by:

i. Number of raters: More raters can increase the ICC value.

ii. Rater expertise: Variability in rater expertise can affect ICC.

iii. Grading scale: The type of scale used (e.g., ordinal, continuous) can impact ICC.

RESULTS AND DISCUSSION

Results of the ICC analysis of Seminar and Project presentations of both Year 2 (ND II) and Year 4 (HND II) students are presented. Five variables (*Physical Appearance, Introduction, Presentation, Practical Application, and Contribution to Knowledge*) were graded for each

student by each rater, and the results are analyzed using both statistics. The average ICC value represents the overall agreement among multiple raters grading the students. It quantifies the consistency and reliability of the ratings. Cronbach's Alpha measures internal consistency, assessing how well scores within a grading measure a single student.

Table 1 shows the average ICC values for the seminar presentations of the Year 2 students for the five variables examined in the two academic sessions among the three raters. The level of agreement among the raters ranges from "*Fair*" to "*Excellent*" (0.593 to 0.973). Generally, there are higher level of agreements in the second session when compared to the first. Also, *Presentation, Introduction, and Physical Appearance* grading are more consistent with good reliability among raters. The variable "*Contribution to Knowledge*" receives the least overall agreement rating and questionable/unacceptable reliability across the two sessions.

Table 1. Average ICC Coefficient for Year 2 Seminar Presentation

Variable	Session	ICC		Cronbach's Alpha	
		Value	Agreement	Value	Reliability
Physical Appearance	First	0.704	Good	0.704	Acceptable
	Second	0.789	Excellent	0.839	Good
Introduction	First	0.677	Good	0.706	Acceptable
	Second	0.840	Excellent	0.872	Good
Presentation	First	0.740	Good	0.769	Acceptable
	Second	0.950	Excellent	0.973	Excellent
Practical Application	First	0.630	Good	0.623	Questionable
	Second	0.560	Fair	0.588	Unacceptable
Contribution to Knowledge	First	0.581	Fair	0.593	Unacceptable
	Second	0.527	Fair	0.605	Questionable

The results of agreements and reliability statistics for the grading of Project for the Year 2 students are presented in Table 2. Only "*Presentation*" and "*Contribution to Knowledge*" for the second and first sessions, respectively, have *good* reliability. Generally, the table shows that there is *poor* internal consistency in the ratings. *Good* consistency/agreement is obtained for only "*Presentation*" in both sessions.

Table 2. Average ICC Coefficient for Year 2 Project Presentation

Variable	Session	ICC		Cronbach's Alpha	
		Value	Agreement	Value	Reliability
Physical Appearance	First	0.506	Fair	0.506	Unacceptable
	Second	0.369	Poor	0.467	Unacceptable

Introduction	First	0.105	Poor	0.103	Unacceptable
	Second	0.355	Poor	0.661	Questionable
Presentation	First	0.678	Good	0.682	Questionable
	Second	0.678	Good	0.841	Good
Practical Application	First	0.464	Fair	0.463	Unacceptable
	Second	0.532	Fair	0.547	Unacceptable
Contribution to Knowledge	First	0.854	Excellent	0.853	Good
	Second	0.276	Poor	0.493	Unacceptable

Table 3 shows both agreements and reliability values for the seminar grading of the Year 4 (HND II) students. There are higher agreements and reliabilities in the ratings in the first session across the five examined variables. This may be largely due to a higher number of raters (3) in the first session compared to those in the second session (2), since more raters can increase both the ICC and the Cronbach's alpha values. *Introduction* and *Presentation* received the most consistent and reliable ratings in the first session considered.

Table 3. Average ICC Coefficient for Year 4 Seminar Presentation

Variable	Session	ICC		Cronbach's Alpha	
		Value	Agreement	Value	Reliability
Physical Appearance	First	0.701	Good	0.773	Acceptable
	Second	0.045	Poor	0.360	Unacceptable
Introduction	First	0.844	Excellent	0.842	Good
	Second	0.076	Poor	0.180	Unacceptable
Presentation	First	0.814	Excellent	0.819	Good
	Second	0.101	Poor	0.097	Unacceptable
Practical Application	First	0.486	Fair	0.491	Unacceptable
	Second	0.083	Poor	0.087	Unacceptable
Contribution to Knowledge	First	0.677	Good	0.702	Acceptable
	Second	0.039	Poor	0.151	Unacceptable

The results of the ratings of the project presentation for the Year 4 students are presented in Table 4. The table reveals that only the *Introduction* in the first session has a *good* internal consistency and an *acceptable* reliability with a Cronbach's alpha value of 0.723. Since four raters were involved the ratings in the first session, while three raters did the second session, both the agreement level and the internal consistency are generally higher in the first session compared to the second session. Like the case in most of the cases examined in

Tables 1 to 3, "*Contribution to Knowledge*" has the least agreement and reliability among the five rated variables.

Table 4. Average ICC Coefficient for Year 4 Project Presentation

Variable	Session	ICC		Cronbach's Alpha	
		Value	Agreement	Value	Reliability
Physical Appearance	First	0.671	Good	0.675	Questionable
	Second	0.220	Poor	0.230	Unacceptable
Introduction	First	0.695	Good	0.723	Acceptable
	Second	0.074	Poor	0.129	Unacceptable
Presentation	First	0.275	Poor	0.416	Unacceptable
	Second	0.302	Poor	0.594	Unacceptable
Practical Application	First	0.459	Fair	0.470	Unacceptable
	Second	0.165	Poor	0.212	Unacceptable
Contribution to Knowledge	First	0.111	Poor	0.333	Unacceptable
	Second	0.015	Poor	0.028	Unacceptable

CONCLUSION

To assess how consistently multiple assessors rated students' performance and to identify areas of disagreement that could compromise the fairness of academic evaluations, this research evaluates inter-rater agreement in the grading of students' seminar and project scores using the ICC. Data on some variables relating to Seminar and Project presentation scores were analyzed, and the results revealed that the levels of agreement among the raters are highly consistent for:

i. "*Presentation*" for the Year 2 (NDII) *Seminar* results for both sessions.
ii. "*Contribution to Knowledge*" for the Year 2 *Project* results for the first session
iii. "*Introduction*" and "*Physical Appearance*" for the Year 2 *Seminar* results for the second session
iv. "*Introduction*" and "*Physical Appearance*" for the Year 4 *Seminar* results for the first session
v. "*Introduction*" and "*Physical Appearance*" for the Year 4 *Project* results for the first session

The findings showed some level of consistency among raters, although improvements can still be made through training and standardized evaluation tools. It can therefore be concluded that the levels of agreement of raters in some variables are generally poor and unacceptable, although some levels of consistency exist for some variables. The

use of ICC revealed moderate to good agreement. These discrepancies underscore the influence of human subjectivity in academic evaluation, particularly in the assessment of inter-rater reliability; biases can persist unnoticed, affecting students' academic outcomes. Thus, the study affirms the necessity of embedding statistical reliability checks within the institutional grading policies to ensure fair, valid and credible academic assessments.

From the findings in the study, the following recommendations are made:

i. Institutionalizing of Training for Raters: Lecturers should undergo periodic rater calibration and training sessions to harmonize grading standards and reduce subjectivity.
ii. Adoption of Standard Grading Rubrics: Institutions should develop and enforce detailed rubrics with objective scoring criteria for every assessment component to minimize interpretation variability among raters.
iii. Moderation Processes: There should be internal moderation where multiple raters or assessors independently grade a subset of assessments before final scores are finalized, and also mandate policies that ensure all subjective assessments are reviewed for inter-rater agreement before final submission.
iv. Digital Assessment Platforms: The adoption of digital tools that integrate rubric-based grading and real-time monitoring of rater behaviour for transparency should be considered.
v. Use of Reliability Statistics: Tools like the ICC should be routinely employed to audit the reliability of scores across assessors or raters, especially in practical or project-based assessments.

REFERENCES

Adebayo, A., & Ediale, T. (2024). Oral defence reliability among polytechnic faculty using ICC. *Journal of Applied Education Studies, 11*(1), 60–71.

Adebayo, F., & Eze, M. (2023). Impact of assessor training on essay scoring reliability. *Journal of Educational Assessment in Nigeria, 7*(3), 44–55.

Adebowale, T., & Ogunyemi, A. (2021). Assessing reliability in clinical evaluations: An ICC approach. *Journal of Educational Measurement in Africa, 9*(2), 112–124.

Adeola, B., & Ajayi, T. (2025). Improving grading reliability through rubric alignment and rater training. *Journal of African Educational Research, 12*(2), 33–48.

Ajibola, S., & Fatimah, K. (2021). Long-term effects of rater workshops in tertiary institutions. *Nigerian Journal of Teacher Education, 9*(1), 78–89.

Akinwunmi, D., & Dada, O. (2021). Standardizing practical exam assessments in universities. *Educational Journal of Practical Pedagogy, 5*(1), 91–102.

Banjo, R., & Etim, U. (2022). Comparing ICC and Kappa for undergraduate thesis evaluation. *African Journal of Educational Statistics, 10*(2), 71–83.

Bello, K., & Uche, O. (2023). Technology-enhanced grading: Improving rater consistency. *Lagos Journal of Education and Technology, 6*(1), 25–37.

Boateng, K., & Amponsah, J. (2020). The effect of digital rubrics on grading efficiency in Ghanaian polytechnics. *West African Journal of Education, 4*(2), 90–101.

Chukwuemeka, C., & Odili, A. (2022). Scoring alignment in business studies assessments. *Journal of Social Science Pedagogy, 8*(4), 55–69.

Ede, O., & Oduwole, T. (2022). Calibration sessions and their effects on ICC scores in Nigerian polytechnics. *Journal of Polytechnic Education, 13*(1), 47–61.

Fatoba, O., & Ajayi, J. (2020). Bias and variability in continuous assessment. *Journal of Education and Social Research, 9*(2), 103–115.

Ibe, R., & Onuoha, S. (2020). Disagreement in practical science grading: A Kappa analysis. *Science Education Review, 10*(1), 36–48.

Jones, D., Kimani, L., & Omondi, P. (2023). Rater background effects on scoring in Kenya's teacher colleges. *East African Education Journal, 11*(2), 62–74.

Kwarteng, K. (2023). A statistical primer on ICC and Kappa for educational research. *Ghanaian Journal of Quantitative Methods, 5*(1), 14–30.

Molefe, T., & Van Rensburg, L. (2021). Assessing final-year project grading reliability using ICC in South Africa. *South African Journal of Higher Education, 35*(2), 130–145.

Musa, Y., & Ibrahim, S. (2024). Fairness in oral presentation scoring using Cohen's Kappa. *Nigerian Journal of Communication and Assessment, 8*(1), 51–64.

Nwankwo, C., & Ahmed, H. (2020). Best practices in inter-rater reliability for tertiary institutions. *Nigerian Journal of Educational Statistics, 6*(2), 70–89.

Okeke, A., & Olumide, R. (2021). Rater training outcomes in essay vs. practical assessments. *Journal of Education and Training Research, 9*(3), 100–115.

Okoro, I., & Eze, U. (2022). Departmental differences in scoring practices in Nigerian colleges. *Journal of Comparative Educational Assessment, 7*(2), 33–47.

Oladimeji, K., & Yusuf, A. (2023). Grading laboratory reports: A Nigerian polytechnic case study using ICC. *West African Journal of Technical Education, 5*(3), 76–88.

Oladipo, M., & Ayeni, R. (2020). Cognitive load and reliability in report grading. *Journal of Educational Psychology and Measurement, 4*(1), 29–40.

Omole, A., & Adebiyi, M. (2023). Disciplinary differences in polytechnic grading. *Nigerian Education and Evaluation Journal, 10*(1), 58–73.

Oseni, O., & Lawal, T. (2024). Electronic scoring sheets and inter-rater agreement in large classes. *Ilorin Journal of Digital Learning, 6*(2), 41–56.

Wang, J., & Chan, P. (2020). Peer presentation evaluation: Comparing ICC and Kappa. *Asian Journal of Educational Research, 11*(4), 101–117.

Wang, X., Harris, B., & Moore, D. (2022). Essay grading agreement among UK educators. *Journal of Educational Evaluation, 9*(3), 55–68.

Yusuf, I. (2023). Common pitfalls in using Kappa and ICC in educational statistics. *Journal of Nigerian Quantitative Research, 7*(2), 77–89.