Benin Journal of Educational Studies
Volume 30, Issue 1&2, July and December; 2024
Official Publication of Institute of Education, University of Benin, Nigeria.
https://beninjes.com/index.php/index

Rater Errors in Teacher Trainee Performance Assessments in a Nigerian University

[1]Dr K. Imasuen and [2]Dr N. Idugboe
[1]Institute of Education, University of Benin, Benin City
[2]Dept. of Educational Evaluation & Counselling Psychology, University of Benin
E-mail: kennedy.imasuen@uniben.edu
Phone Number: +2348058824653

Abstract: *This study investigated the prevalence of rating errors in teacher trainee performance assessments within a Nigerian university, focusing on how evaluators applied the Teaching Practice Performance Assessment Scale (TPPAS). Teaching practice is a compulsory component of teacher education, yet its reliability is often undermined by subjectivity in rater judgments. Five raters who supervised postgraduate diploma trainees during their practicum were purposively selected, and their completed rating sheets were analyzed. The TPPAS, a validated instrument with a Cronbach's alpha of 0.87, was used to capture multiple dimensions of teaching competence. Data analysis employed descriptive statistics alongside error-detection metrics using the R software, which allowed raters to be classified according to severity, leniency, central tendency, and halo effect biases. The results revealed that all raters demonstrated at least one form of error, with the halo effect emerging as the most widespread. Leniency bias was observed among three raters, while severity bias appeared in two, and central tendency was particularly evident in those who clustered scores around the midpoint of the scale. These findings confirm that subjectivity is systemic in teacher trainee evaluations and raise concerns about the fairness and validity of current practices. The study concludes that such biases compromise the credibility of teaching practice assessments and recommends the adoption of structured rater training, use of standardized rubrics, multiple raters, and regular monitoring systems to strengthen the reliability of teacher education assessments.*

Keywords: *Performance assessment, Rater bias, Rating errors, Halo effect*

Introduction

Education globally, including in Nigeria, is widely recognized as a major driver of societal transformation. This transformation occurs through learning, which is often defined as a change in behaviour or competence following engagement in educational processes. For learning to achieve its intended objectives, teachers play a central role. They not only transmit knowledge but also evaluate its assimilation and make decisions about learner progression and certification (Imasuen & Aibinuomo, 2022; Babatimehin et al., 2025). Teachers further assess whether the broader educational goals have been realized, highlighting their importance in ensuring quality education.

How to Cite

Imasuen , K., & Idugboe, N. (2025). Rater Errors in Teacher Trainee Performance Assessments in a Nigerian University. Benin Journal of Educational Studies, 30(1&2), 43–49. Retrieved from https://beninjes.com/index.php/bjes/article/view/150

To function effectively, teachers require comprehensive training that integrates pedagogy, assessment techniques, and classroom management skills. In teacher education institutions, after completing coursework, trainee teachers must demonstrate their competencies during teaching practice or performance assessment. This component is compulsory and a prerequisite for recognition by the Teachers Registration Council of Nigeria (TRCN). Depending on institutional policies, this practicum typically lasts six to eight weeks (Omodan, 2023).

According to Bloom's Taxonomy, cognitive skills are categorized into lower-order (e.g., Knowledge, remembering, understanding) and higher-order skills (analyzing, evaluating, synthesis). While lower-order skills are frequently assessed through traditional testing methods, higher-order skills demand more complex, process-oriented evaluations that integrate cognitive, affective, and psychomotor domains (Kutlu et al., 2014). Performance assessments are especially suitable for capturing higher-order competencies because they emphasize authentic, real-world tasks and require critical thinking and application (Şata & Karakaya, 2022; Brown & Hudson, 2022).

Despite their value, performance assessments face challenges in ensuring objectivity. Unlike standardized tests, performance ratings are susceptible to variability among evaluators. Recent studies confirms that rater effects such as halo, leniency, severity, and central tendency biases distort the reliability and validity of performance-based judgments (Michela, 2022; Schmidt et la., 2023). Approaches such as rubric development, rater training, multiple raters, and automated scoring have been proposed to mitigate these challenges (Rodríguez et al., 2023; Şata & Karakaya, 2022).

A persistent challenge, however, lies in achieving consistency among raters. Even with standardized criteria, raters may interpret and apply them differently, resulting in variability that reflects subjective tendencies rather than actual trainee performance. This phenomenon, referred to as rater effects, introduces unwanted variance unrelated to the trainee's competence, thereby threatening score validity (Liu & Zhang, 2022; Abe, 2022). Common errors include halo effect, where a general impression of a trainee influences judgments across unrelated domains (Michela, 2022; Schmidt et al., 2023); leniency bias, where raters assign overly generous scores; severity bias, where raters are consistently more critical; and central tendency bias, where raters avoid extreme ratings and cluster around the midpoint (Şata & Karakaya, 2022). These biases reduce the accuracy of performance assessments, raising concerns about fairness and credibility.

Recent studies stress the importance of addressing rater errors. Rodríguez et al. (2023) found that targeted rater training significantly reduced halo and central tendency biases in teacher evaluations. Similarly, Liu and Zhang (2022) emphasized the role of cultural factors in shaping rater perceptions, highlighting the need for context-sensitive training. Nigerian-based research also confirms that teachers often display leniency and severity biases in grading (Abe, 2022), and many lack sufficient training in school-based assessment practices (Babatimehin et al., 2025).

Statement of the Problem

Performance assessments are a cornerstone of teacher education, providing evidence of trainees' readiness for professional practice. However, the reliability and validity of these assessments can be significantly undermined by rater biases, including halo effect, leniency, severity, and central tendency errors (Michela, 2022; Rodríguez et al., 2023). These errors distort assessment outcomes and compromise the goals of teacher training institutions. Despite growing recognition of these biases, empirical research exploring their prevalence and impact in Nigerian teacher trainee evaluations remains limited. Raters, who serve as gatekeepers for teacher certification, may bring subjective tendencies into the evaluation process, leading to inaccurate and unfair ratings. Consequently, there is

an urgent need to examine the nature and extent of rating errors in the Nigerian context to inform strategies for improving the fairness and credibility of teacher education assessments.

Research Questions

The following questions guided the study

1. What are the prevalent rating errors among raters evaluating teacher trainees?
2. How does the severity of ratings vary among supervisors?

Methodology

This study employed a descriptive survey design. The population consisted of raters who supervised and evaluated teacher trainees undertaking the Postgraduate Diploma in Education (PGDE) practicum at a Nigerian university. From this population, a purposive sample of five raters was selected, as these individuals were directly responsible for assessing trainee teachers during the teaching practice exercise. The instrument used for data collection was the Teaching Practice Performance Assessment Scale (TPPAS), a structured five-point Likert scale designed to evaluate multiple dimensions of teaching performance. These dimensions included lesson planning, instructional delivery, classroom management, and communication skills, with the scale ranging from *Poor* (1) to Excellent (5). The validity of the TPPAS was established through expert judgment. Experts in educational measurement and teacher educators reviewed the instrument to ensure that the items were representative of teaching competencies and aligned with the professional standards of the Teachers Registration Council of Nigeria (TRCN). Their input helped refine the instrument, thereby strengthening its content validity. Reliability was confirmed through a pilot test conducted with raters not included in the main study. The internal consistency of the scale was measured using Cronbach's alpha, which produced a coefficient of 0.87, demonstrating a high level of reliability. Inter-rater reliability was also examined by comparing the consistency of ratings across different evaluators, and the results showed an acceptable level of agreement, indicating that the TPPAS was a dependable measure of teaching performance.

In order to classify rating errors, specific criteria were applied. Severity and leniency were determined by calculating severity error z-scores and raw deviations from the overall group mean. Negative z-scores indicated leniency, meaning the rater tended to score trainees higher than average, while positive z-scores indicated severity, reflecting a tendency to score more critically. The magnitude of the z-score showed the extent of each bias. Central tendency error was identified through within-rater standard deviation, variance, and the proportion of mid-score ratings. Raters with low variability in their ratings and frequent use of the midpoint score were considered to demonstrate central tendency. Finally, the halo effect was assessed by examining average inter-item correlations. Very high correlations, at or above 0.90, indicated that a rater's scores were influenced by an overall impression rather than by differentiated judgments of specific competencies.

Data collection involved retrieving completed rating sheets after the teaching practice exercise. Each rater's scores were analyzed according to the classification criteria described above, and data analysis was conducted using the R software to calculate descriptive statistics such as means, standard deviations, and percentages. These were complemented by the error-detection indices, which allowed raters to be categorized according to their biases. The findings were presented both in prose and in tabular form to clearly highlight the differences among raters.

Results

Table 1

Summary of Rating Errors among Supervisors Evaluating Teacher Trainees

Raters	Severity (Z Score)	Severity (Raw Deviation)	Central Tendency (Standard deviation)	Mid-Score Rate	Leniency Mean	Halo (Average Inter-Item Correlation)	Dominant Errors Identified
1	-0.34	-0.28	0.69	0.02	2.64	0.84	Primarily Leniency
2	-0.68	-0.58	0.55	0.00	2.35	0.91	Leniency and Halo Effect
3	0.65	0.54	0.48	0.27	3.47	0.96	Severity, Central Tendency and Halo Effect
4	-0.81	-0.68	0.52	0.01	2.25	0.91	Severity and Halo Effect
5	0.37	0.31	0.58	0.02	3.24	0.92	Severity, Halo Effect, and Central Tendency

The analysis in Table 1 reveals that all raters demonstrated one or more forms of rating error, indicating that rating biases are prevalent in the evaluation of teacher trainees. Leniency Error was observed in multiple raters. Rater 1 and Rater 2 both displayed this tendency, with negative severity Z-scores (-0.34 and -0.69, respectively). Rater 4 recorded the most pronounced leniency, with the most negative severity Z-score (-0.81). These results indicate that a significant proportion of raters consistently rated teacher trainees higher than the group average, suggesting a tendency toward generosity in their assessments.

Severity error was also evident among the raters. Rater 3 showed the strongest severity bias, with a positive severity Z-score of 0.65, making them the most critical rater in the group. Rater 5 also leaned toward severity, with a positive Z-score of 0.37. This indicates that some raters consistently rated trainees below the overall mean, suggesting a harsher evaluation style. Central Tendency Error was detected in Raters 3 and 4. Rater 3, in particular, showed the strongest central tendency bias, with the lowest within-rater standard deviation (0.48) and the highest mid-score rate (0.27). This reflects an avoidance of extreme ratings and a clustering of scores around the middle category, thereby failing to capture performance differences effectively. Rater 4 also showed a moderate central tendency error, with a low within-rater standard deviation (0.52), further reinforcing this pattern.

Halo effect error emerged as the most pervasive error across raters. Raters 2, 3, 4, and 5 all demonstrated very high inter-item correlations (ranging from 0.91 to 0.96), indicating that their ratings were heavily influenced by a general impression rather than by differentiated judgments of specific competencies such as lesson planning, instructional delivery, or communication skills. Rater 3, with the highest correlation (0.96), exhibited the strongest halo effect error, effectively assigning nearly uniform ratings across all items.

In all, the prevalent rating errors among the raters were leniency, severity, central tendency, and halo effect errors, with halo bias being the most widespread. Leniency was more common among Raters 1, 2, and 4; severity was notable for raters 3 and 5, while raters 3 and 4 also demonstrated central tendency errors. The consistent presence of halo error across nearly all raters highlights a systemic issue in rating practices, suggesting that raters often fail to differentiate between various dimensions of trainee performance.

Table 2

Variation in Severity and Leniency of Supervisors' Ratings

Raters	Mean Score (out of 5)	% Excellent Ratings (Rate 5)	Severity/Leniency Indication
1.0	4.29	61.11%	Strongest Leniency (Most Generous)
2.0	3.92	41.67%	Moderate Leniency
2.5	3.79	25.00%	Strongest Severity (Most Conservative)
3.0	3.87	33.33%	Moderate Severity
4.0	3.96	38.89%	Mild Severity

In Table 2, the severity and leniency of raters' ratings vary considerably, as shown by differences in mean scores and the proportion of "Excellent" ratings awarded. Rater 1 emerges as the most lenient, with a high mean score of 4.29 (out of 5) and 61.11% of ratings falling in the "Excellent" category. In contrast, Rater 3 is the most severe, with a lower mean score of 3.79 and only 25% of ratings classified as "Excellent." The difference between these two extremes is notable. The mean scores differ by 0.50 points, while the proportion of "Excellent" ratings differs by 36.11 percentage points. Such large variations indicate that raters apply rating standards inconsistently, with some being overly generous and others unduly conservative. This inconsistency presents a lack of standardization in applying the Teaching Practice Performance Assessment Scale (TPPAS). Instead of reflecting trainee competence alone, the ratings are heavily influenced by the individual rater's severity or leniency bias. Consequently, trainee outcomes may be more reflective of who assessed them rather than their actual performance

Discussion of Findings

The findings of this study reveal that raters evaluating teacher trainees exhibited multiple forms of rating errors, including leniency, severity, central tendency, and halo effect errors. These biases varied in type and intensity across raters, influencing the reliability and fairness of the Teaching Practice Performance Assessment Scale (TPPAS). The most prevalent error identified was the halo effect, observed in Raters 2, 3, 4, and 5. This indicates that raters often relied on a global impression of trainees rather than differentiating among specific performance dimensions such as lesson planning, instructional delivery, or classroom management. Michela (2022) emphasizes that halo bias is particularly pervasive in educational evaluations because raters rely on cognitive shortcuts, while Schmidt et al.et al. (2023) demonstrate experimentally that halo distortions can carry over from one task to another even when performance is unrelated. These findings affirm that halo tendencies compromise the objectivity of teacher trainee assessments, making it difficult to capture accurate profiles of trainee strengths and weaknesses.

Leniency and severity errors also featured prominently. Raters 1, 2, and 4 demonstrated leniency, while Raters 3 and 5 showed severity, with Rater 3 emerging as the most severe. This variability highlights that rater stringency is not standardized, reflecting personal dispositions and subjective benchmarks. Abe (2022), in a Nigerian context, similarly found that teachers exhibited leniency and severity biases in assessing secondary school students, leading to inconsistent grading. Internationally, Babatimehin et al. (2025) showed that many teachers lack sufficient knowledge of assessment practices, which exacerbates rating inconsistencies. Such evidence confirms that trainees' scores may be shaped as much by who assessed them as by their actual teaching performance.

Central tendency error was most evident among Raters 3 and 4, particularly Rater 3, who avoided extreme ratings and clustered scores around the mid-point. This practice conceals genuine performance differences and undermines the diagnostic value of assessments. Şata and Karakaya (2022) reported similar findings, noting that raters uncertain about applying criteria often "play safe" by selecting mid-scale ratings, which reduces reliability. Further, this study revealed wide variations in severity levels among raters. Rater 1, the most lenient, recorded a mean score of 4.29 out of 5, with 61.11% of ratings in the "Excellent" category, while Rater 3, the most severe,

recorded a mean of 3.79, with only 25% "Excellent" ratings. The 0.50-point difference in mean ratings and a 36% gap in Excellent ratings highlight the lack of rating consistency. Recent scholarship has underscored similar concerns. Rodríguez et al. (2023) argue that such differences represent systematic measurement errors, while Liu and Zhang (2022) emphasize that cultural and contextual factors further shape raters' judgmental patterns.

Overall, these findings confirm that rater effects are systemic in teacher trainee evaluations. Every rater in this study demonstrated one or more types of bias, with the halo effect emerging as the most widespread. These results are consistent with global and Nigerian evidence showing that subjectivity in rater practices undermines the validity of performance-based assessments (Abe, 2022; Michela, 2022; Schmidt et al., 2023; Babatimehin et al., 2025). Unless addressed, such inconsistencies threaten the credibility of teaching practice assessments, disadvantaging some trainees while inflating the performance of others.

Conclusion

Based on the study's findings, it was concluded that all raters exhibited one or more forms of bias. The most prevalent error was the halo effect, followed by leniency, severity, and central tendency errors. Rater 1 was the most lenient, while Rater 3 was the most severe, demonstrating the lack of consistency across evaluators. Overall, the results point to the systemic nature of rater effects in teacher trainee evaluations. Without corrective measures, teaching practice scores risk reflecting rater bias rather than trainee competence. This threatens the fairness, reliability, and credibility of teacher education in Nigeria and beyond.

Recommendations

Based on the findings, the following recommendations are made:

1. Teacher education institutions should provide continuous training for raters to raise awareness of rating errors and strategies for minimizing them.
2. Clear and well-structured rubrics should be developed and applied consistently across all raters to ensure that judgments are tied to specific, observable performance indicators
3. Each trainee should be evaluated by more than one rater, as this practice reduces the impact of individual biases and increases the reliability of scores.
4. Institutions should establish systems to regularly review rater performance, provide feedback, and detect systematic leniency, severity, or halo patterns among raters.
5. Rater training programmes should consider cultural influences on judgment to promote fairness and inclusivity in evaluations.

References

Abe, T. O. (2022). Teachers' bias in assessment process in selected school subjects in Ekiti State, Nigeria. *Journal of Research in Education and Society (JRES)*, 13(1), 27–37. https://icidr.org.ng/index.php/Jres/article/view/1115

Babatimehin, O., Adebisi, A., & Adeyanju, O. (2025). Assessing teachers' knowledge of school-based assessment practices in Nigerian secondary schools. *Discover Education*, 4, 41. https://doi.org/10.1007/s44217-025-00512-8

Brown, J. D., & Hudson, T. (2022). *Assessing language performance: Issues and practices*. Routledge.

Imasuen, K., & Aibinuomo, O. (2022). Teachers' role in the realization of educational objectives in Nigeria. *African Journal of Educational Studies*, 19(2), 101–115.

Kutlu, Ö., Doğan, C. D., & Karakaya, İ. (2014). *Performance-based assessment: From theory to practice*. Ani Publishing.

Liu, J., & Zhang, Y. (2022). Cultural influences on rater bias in performance assessment: A systematic review. *Assessment in Education: Principles, Policy & Practice*, 29(4), 421–438. https://doi.org/10.1080/0969594X.2021.2003205

Michela, E. (2022). Toward understanding and quantifying halo in students' evaluations of teaching. *Assessment & Evaluation in Higher Education*, 47(7), 1061–1075. https://doi.org/10.1080/02602938.2022.2086965

Omodan, B. I. (2023). Teacher training, professional practice, and the challenges of quality education in Nigeria. *Journal of Education, Society and Behavioural Science*, 36(2), 1–11. https://doi.org/10.9734/jesbs/2023/v36i2825

Rodríguez, J., Pérez, M., & García, R. (2023). Reducing halo and central tendency bias in teacher evaluations: The role of rater training. *Educational Assessment, Evaluation and Accountability*, 35(2), 129–149. https://doi.org/10.1007/s11092-023-09427-9

Şata, M., & Karakaya, İ. (2022). Investigating the impact of rater training on rater errors in the process of assessing writing skills. *International Journal of Assessment Tools in Education (IJATE)*, 9(3), 492–508. https://doi.org/10.21449/ijate.877035

Schmidt, F. T. C., Kaiser, J., & Retelsdorf, J. (2023). Halo effects in grading: An experimental approach. *Journal of Educational Psychology*, 115(4), 665–678. https://doi.org/10.1037/edu0000761