AJSTME, Volume. 12, Issue 2. 230-240, March, 2026. https://www.ajstme.com.ng
ISSN: 2251-0141 (Print), 2971-6233 (Online)

Comparing the Reliability of Economics Essay Scores Under Human and AI-Assisted Scoring: A Generalizability Theory Analysis

Ossai Elizabeth Ngozika
Department of Science Education, University of Nigeria, Nsukka
Email: elizabethngozika@gmail.com

Abstract

This study compared the reliability of Economics essay scores under human-rater and Artificial Intelligence (AI)-assisted scoring conditions using Generalizability Theory (G-Theory). Four research questions and two hypotheses guided the study using a fully crossed p × i × r design (Persons × Items × Raters). The population comprised 318 Senior Secondary School II (SS II) Economics students from 15 public secondary schools in Uzo-Uwani Local Government Area, Nsukka Education Zone, Enugu State, Nigeria. Using simple random sampling, 50 students were selected for the study. The Economics Essay Test (EET) was developed and validated by experts in Economics Education and Measurement and Evaluation. Five essay items were scored independently by two human raters and ChatGPT (GPT-5.5, OpenAI) using a standardized scoring rubric. Data were analyzed using the gtheoryr and lme4 packages in RStudio. Variance components were estimated to identify sources of measurement error, while hypotheses were tested at the 0.05 level of significance. The findings revealed that measurement error constituted the largest variance component, followed by item effects, whereas student variance was minimal. A statistically significant difference was found between human-rater and AI-assisted scoring conditions, $F(2, 698) = 35.742$, $p < .001$, leading to the rejection of the null hypothesis. The generalizability coefficient ($E\rho^2 = 0.2773$) and dependability coefficient ($\Phi = 0.2773$) indicated low reliability of the Economics Essay Test scores. The study concluded that the Economics Essay Test produced low reliability and limited score generalizability under both human-rater and AI-assisted scoring conditions. It was recommended that the quality and number of essay items be improved and that a calibrated hybrid human–AI-assisted scoring approach be adopted to enhance the reliability and consistency of Economics essay test scores.

Keywords: AI-Assisted Scoring, Economics Essay Assessment, Educational Measurement, Generalizability Theory, Human Scoring, Measurement Error, Reliability.

Introduction

Reliable assessment is fundamental to educational decision-making because it ensures that students' scores accurately reflect their knowledge, skills, and abilities rather than unintended sources of measurement error. In Economics education, assessment provides evidence of students' understanding of economic concepts, ability to analyse economic issues, apply theoretical principles, and communicate economic reasoning. Consequently, the reliability of assessment scores is essential, particularly when such scores are used for certification, placement, and admission into higher institutions. Economics essay assessment is an important component of educational evaluation because it enables students to demonstrate higher-order cognitive skills such as explanation, analysis, evaluation, problem-solving, and the application of economic principles to real-life situations. Essay assessment requires students to organise ideas logically, justify conclusions with evidence, and communicate economic reasoning effectively. However, because essay responses require human judgment during scoring, they are susceptible to variations in rater severity, interpretation of scoring criteria, fatigue, and inconsistency in judgment, all of which contribute to measurement error and reduce score reliability (Brennan, 2001; Engelhard, 2013). The issue of reliable essay scoring is particularly important within the Nigerian secondary school education system. Economics is one of the core subjects offered at the senior secondary school level, and students' ability to explain economic concepts, analyse economic issues, and apply economic principles is assessed through essay questions in the

West African Senior School Certificate Examination (WASSCE) conducted by the West African Examinations Council (WAEC). Because the results of this examination are used for certification and admission into tertiary institutions, ensuring that essay scores are reliable, consistent, and fair is essential for valid educational decision-making.

Traditionally, Economics essay responses have been scored by trained human raters using standardized marking schemes and scoring rubrics. Human scoring offers important advantages, including professional judgment, contextual understanding, and the ability to recognise different but equally valid approaches to solving economic problems. Nevertheless, research has shown that human scoring may be influenced by differences in rater severity, inconsistent interpretation of scoring criteria, scorer fatigue, and subjective judgment, thereby introducing measurement error into assessment outcomes (Engelhard, 2013; Williamson et al., 2020). These challenges have stimulated increasing interest in the application of Artificial Intelligence (AI) to support educational assessment.

Recent advances in Artificial Intelligence (AI), particularly large language models (LLMs), have expanded opportunities for educational assessment. AI-assisted scoring systems are increasingly being explored for evaluating written responses, providing feedback, and supporting scoring decisions. Through advances in natural language processing and machine learning, AI systems can analyse large volumes of written responses rapidly and apply scoring criteria with increasing consistency, making them attractive for large-scale educational assessment (Li & Ng, 2024; Misgna et al., 2025).

Despite these promising developments, AI-assisted scoring should not be regarded as an automatic replacement for human judgment. Although AI systems have the potential to improve scoring efficiency and consistency, they may also introduce additional sources of measurement error, including algorithmic bias, prompt sensitivity, differences across AI model versions, limited transparency in scoring decisions, and difficulties in recognising complex disciplinary reasoning. In Economics essay assessment, students may present different but equally valid explanations of economic concepts supported by diverse economic theories or contextual examples. AI systems may therefore experience challenges in recognising nuanced reasoning, contextual interpretations, and culturally embedded expressions. Consequently, the reliability and fairness of AI-assisted scoring require rigorous empirical evaluation before such systems can be confidently adopted for high-stakes educational assessment. Previous studies comparing AI-assisted scoring with human scoring have produced encouraging but mixed findings. Recent research suggests that AI-assisted scoring systems can achieve substantial agreement with human raters and improve scoring efficiency in evaluating written responses. However, their performance varies across assessment contexts, essay characteristics, scoring rubrics, and AI models, indicating that AI-generated scores require careful validation before they are adopted for high-stakes educational assessment (Li & Ng, 2024; Misgna et al., 2025). Moreover, most previous studies have evaluated AI-assisted scoring using correlation coefficients, percentage agreement, or inter-rater reliability statistics. Although these approaches provide useful information regarding the level of agreement between human and AI scoring, they do not adequately identify the multiple sources of measurement error that contribute to score variability in essay assessment. Generalizability Theory (G-Theory), developed by Cronbach and colleagues and subsequently advanced by Brennan (2001), provides a comprehensive framework for evaluating reliability in complex assessment situations.

Unlike Classical Test Theory, which considers measurement error as a single undifferentiated component, Generalizability Theory decomposes observed score variance into multiple facets, including persons, items, raters, and their interactions (Cronbach et al., 1972; Brennan, 2001). Through a Generalizability Study (G-study), variance components associated with different measurement facets are estimated, while a Decision Study (D-study) examines how modifications to the assessment design influence score reliability and dependability. Consequently, Generalizability Theory provides richer information for improving assessment quality than conventional reliability coefficients. The theory is particularly appropriate for the present study because it provides a framework for examining the major sources of variability associated with Economics essay assessment. In this study, students constitute the persons (p) facet, the Economics essay questions represent the items (i) facet, and the two human raters together with the AI-assisted scoring system represent the raters (r) facet. The fully crossed Persons × Items × Raters (p × i × r) design enables the simultaneous estimation of variance attributable to students, essay items, raters, and their interaction effects. Consequently, the design provides a comprehensive evaluation of whether observed differences in Economics essay scores are attributable to students' performance, characteristics of the essay questions, differences between human and AI-

assisted scoring, or interactions among these measurement facets.

Despite the growing application of Artificial Intelligence in educational assessment, empirical evidence regarding the reliability of AI-assisted scoring in Economics essay assessment remains limited. Most previous studies have evaluated AI-assisted scoring using correlation coefficients, percentage agreement, or inter-rater reliability statistics rather than Generalizability Theory, which provides a more comprehensive evaluation of multiple sources of measurement error. Furthermore, few studies have investigated AI-assisted scoring within the Nigerian secondary school context. Consequently, there is insufficient evidence regarding the dependability of Economics essay scores under human and AI-assisted scoring conditions. Therefore, this study applied Generalizability Theory to compare the reliability of Economics essay scores under human and AI-assisted scoring conditions and to provide empirical evidence that can guide the appropriate integration of AI-assisted scoring into educational assessment.

Economics examinations often use essay and other constructed-response items to assess students' higher-order thinking skills. However, scoring such responses is inherently subjective and may be affected by differences in rater judgment, interpretation of scoring rubrics, and scoring conditions, leading to concerns about score reliability and consistency. Recent advances in Artificial Intelligence (AI) have introduced automated scoring systems capable of evaluating written responses. Although these systems show promise, there is limited empirical evidence regarding the reliability of AI-generated scores in Economics assessments. Most previous studies have focused on the level of agreement between human and AI raters, with little attention given to the various sources of measurement error that may affect score dependability. Furthermore, traditional reliability approaches are limited in their ability to examine multiple sources of error simultaneously. As a result, it remains unclear whether AI-assisted scoring improves score reliability or introduces additional sources of measurement error. Therefore, there is a need to investigate the reliability and generalizability of Economics essay scores generated by human and AI raters. This study addresses this gap by applying Generalizability Theory to examine the contributions of students, items, and scoring methods to score variability and to determine the dependability of the resulting scores

AJSTME, Volume. 12, Issue 2. 230-240, March, 2026. https://www.ajstme.com.ng
ISSN: 2251-0141 (Print), 2971-6233 (Online)

Theoretical Framework Diagram (G-Theory Based)

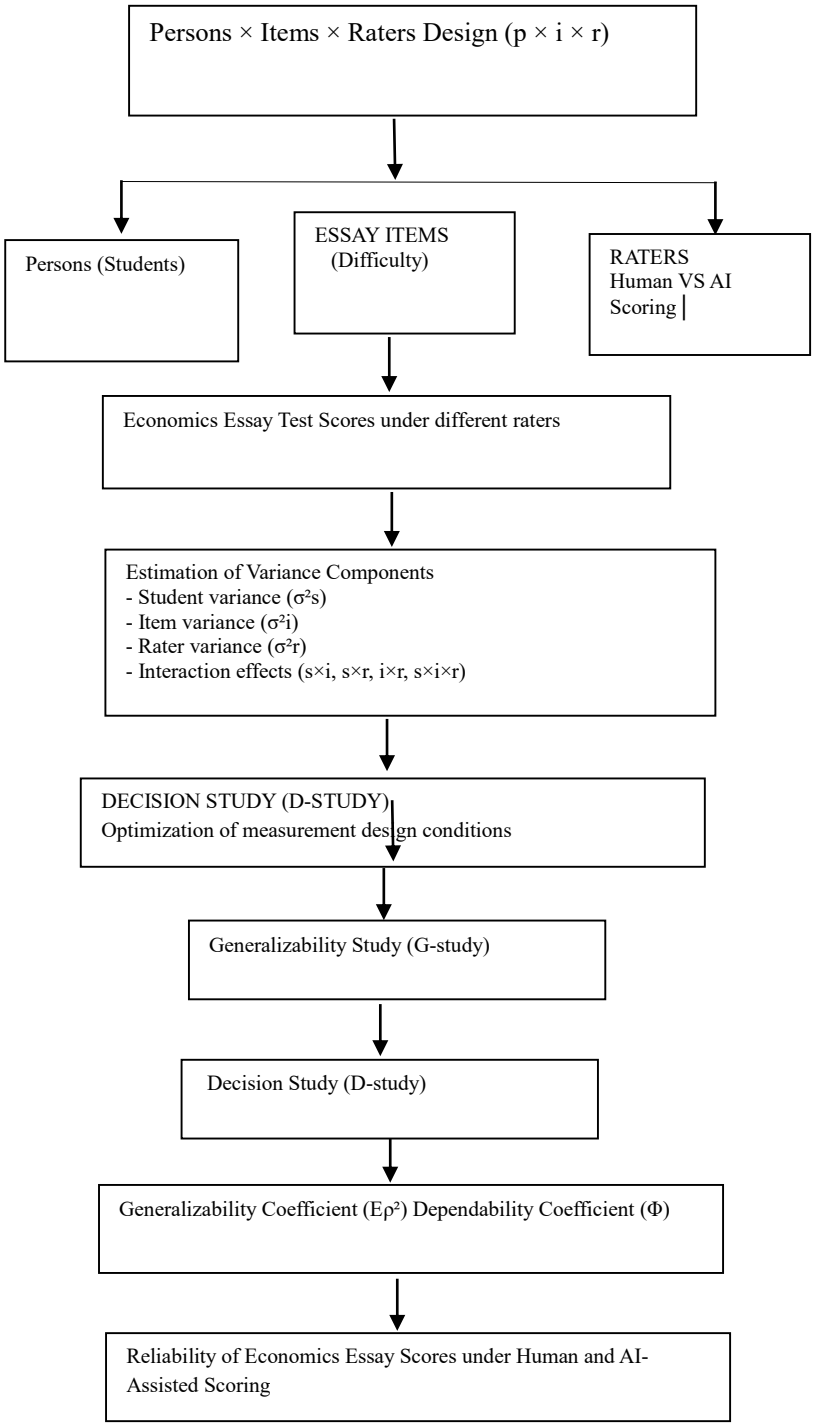

Figure 1

Purpose of the Study

The purpose of this study was to compare the reliability of Economics essay test scores under human-rater and Artificial Intelligence (AI)-assisted scoring conditions using Generalizability Theory.

Specifically, the study seeks to:

1. Determine the extent to which students, essay items, scoring conditions (human raters and Artificial Intelligence), and their interactions contribute to variation in Economics essay test scores.
2. Investigate the difference in Economics essay test scores between human-rater and Artificial Intelligence (AI)-assisted scoring conditions.
3. Estimate the generalizability ($E\rho^2$) and dependability (Φ) coefficients of Economics essay test scores under human-rater and AI-assisted scoring conditions.
4. Examine the level of consistency between human-rater scoring and Artificial Intelligence (AI)-assisted scoring in evaluating Economics essay test responses.

To accomplish the objectives of this study, the following research questions were formulated to guide the investigation:

1. To what extent do students, essay items, scoring conditions (human raters and Artificial Intelligence), and their interactions contribute to variation in Economics essay test scores?
2. What is the difference in Economics essay test scores between human-rater and Artificial Intelligence (AI)-assisted scoring conditions?
3. What are the generalizability ($E\rho^2$) and dependability (Φ) coefficients of Economics essay test scores under human-rater and AI-assisted scoring conditions?
4. How consistent are human-rater scoring and Artificial Intelligence (AI)-assisted scoring in evaluating Economics essay test responses?

Hypotheses

H_{01}: There is no significant contribution of students, essay items, scoring conditions (human raters and Artificial Intelligence (AI)-assisted scoring), and their interactions to the variance in Economics essay test scores.

H_{02}: There is no significant difference in Economics essay test scores between human-rater and Artificial Intelligence (AI)-assisted scoring conditions.

Methodology

Research Design

The study adopted a measurement research design grounded in Generalizability Theory (G-Theory) to evaluate the dependability of Economics essay test scores and identify sources of measurement error. A fully crossed p × i × r design (persons × items × raters) was employed, where all students responded to all essay items and all responses were scored by the two human raters and one AI-assisted scoring system. The design enabled the estimation of variance components associated with persons, items, raters, and their interactions through Generalizability Study (G-study) and Decision Study (D-study) analyses, thereby providing evidence on the reliability and dependability of scores under human and AI-assisted scoring conditions.

Venn diagram showing the relationship among the facets of interest in the generalizability theory study

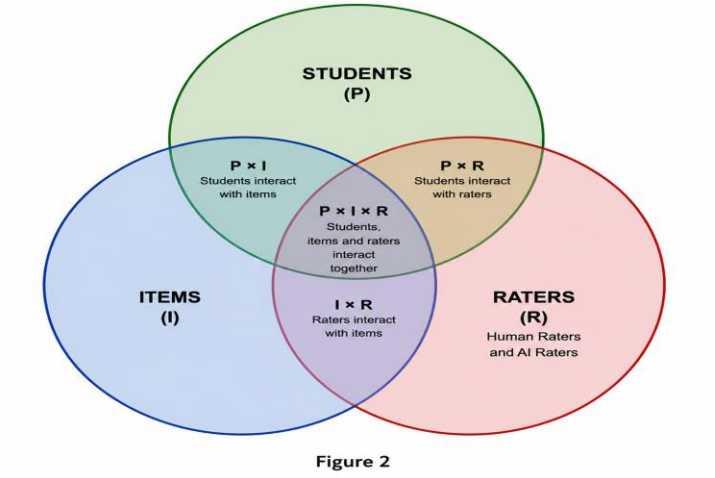

Figure 2

The study was carried out in Uzo-Uwani Local Government Area (LGA) of Nsukka Education Zone, Enugu State, Nigeria. The area comprises both public and private secondary schools where Economics is taught at the Senior Secondary School level. The schools in the area follow the Nigerian secondary school curriculum, and SS II students constitute an important group for assessing Economics learning outcomes. The area was selected because it provides access to the target population and offers a suitable context for investigating the reliability of Economics essay test scores under human-rater and Artificial Intelligence (AI)-assisted scoring conditions.

The population of the study comprised 318 Senior Secondary School II (SS II) Economics students enrolled in 15 public secondary schools in Uzo-Uwani Local Government Area (LGA) of the Nsukka Education Zone, Enugu State. This population figure was obtained from the Planning, Research and Statistics (PRS) Department, Post Primary School Management Board (PPSMB), Nsukka (2025).

The sample for the study consisted of 50 SS2 Economics students drawn from the population of 318 students in Uzo-Uwani Local Government Area (LGA) of Nsukka Education Zone, Enugu State. A simple random sampling technique was used to select the respondents. This was achieved by assigning identification numbers to all 318 students and writing the numbers on identical slips of paper. The slips were thoroughly mixed in a container, and 50 slips were randomly selected without replacement. This procedure ensured that each student had an equal chance of being selected for the study and minimized selection bias.

The instrument used for data collection was the Economics Essay Test (EET) developed by the researcher. The test covered five topics in the SS II Economics curriculum: Demand and Supply, Production, National Income, Market Structures, and Economic Development. These topics were selected because they are core areas of the curriculum and require students to demonstrate understanding and application of economic concepts through essay responses. The EET consisted of five essay questions, each carrying four marks, giving a total score of 20 marks. A scoring rubric developed by the researcher was used by two human raters and an Artificial Intelligence (AI) system to score the students' responses.

The instruments were subjected to face and content validation by three experts, comprising two experts in Measurement and Evaluation and one expert in Economics Education, University of Nigeria, Nsukka, to establish their face and content validity. The experts assessed the relevance, clarity, and adequacy of the items, and their suggestions were incorporated into the final versions of the instruments.

The reliability of the Economics Essay Test (EET) was established using Generalizability Theory (G-Theory). Unlike Classical Test Theory (CTT), which provides a single reliability estimate such as Cronbach's alpha, G-Theory examines multiple sources of measurement error through Generalizability (G-) and Decision (D-) studies. A fully crossed person × item × rater (p × i × r) design was used to estimate the generalizability coefficient ($E\rho^2$) and dependability coefficient (Φ), which assessed the consistency and dependability of scores across students, items, and scoring conditions.

Data were collected by administering the Economics Essay Test (EET) to 50 SS II Economics students under standardized conditions. The scripts were coded, anonymized, and independently scored by two trained human raters and an AI scoring system using the same validated rubric. The resulting scores were recorded for analysis.

Data were analyzed using the gtheoryr and lme4 packages in RStudio. A Generalizability Study (G-study) was conducted to estimate variance components due to students, items, scoring conditions, and their interactions, while a Decision Study (D-study) was used to estimate the generalizability coefficient ($E\rho^2$) and dependability coefficient (Φ). Linear mixed-effects models were fitted to examine differences between human and AI-assisted scoring conditions. Reliability and dependability were considered acceptable when $E\rho^2$ and Φ values were 0.70 or above.

Results

Table 1: Variance Components Attributable to Students, Essay Items, and Residual Error in Economics Essay Test Scores under Human-Rater and Artificial Intelligence (AI)-Assisted Scoring Conditions.

Source of Variance	Variance Component (σ^2)	Contribution to Total Variance (%)
Students (ID)	0.01340	2.49
Essay Items	0.04044	7.53
Residual Error	0.48318	89.98
Total	**0.53702**	**100.00**

Table 1 shows the variance components attributable to students, essay items, and residual error in Economics essay test scores under human-rater and Artificial Intelligence (AI)-assisted scoring conditions. The findings reveal that residual error had the largest variance component (0.48318), accounting for 89.98% of the total variance. This indicates that most of the variability in the scores was attributable to unexplained sources not explicitly modeled in the analysis. Essay items had a variance component of 0.04044, contributing 7.53% of the total variance, suggesting some differences in the functioning or difficulty of the essay questions. In contrast, students had the smallest variance component (0.01340), representing 2.49% of the total variance, indicating limited variability in students' performance. Overall, the findings suggest that the assessment design under human-rater and Artificial Intelligence (AI)-assisted scoring conditions was influenced predominantly by unexplained variability, which may have reduced the reliability of the Economics essay test scores.

Table 2: Comparison of Economics Essay Test Scores under Human-Rater and AI-Assisted Scoring Conditions

Scoring Condition	Estimate (β)	Std. Error	t-value	p-value
AI-Assisted Scoring (Reference Category)	3.572	0.0467	76.462	< .001
Human Rater 1	-0.452	0.0617	-7.322	< .001
Human Rater 2	-0.452	0.0617	-7.322	< .001

Overall Effect of Scoring Condition

Effect	F-value	df	p-value
Scoring Condition (Human Raters and AI-Assisted Scoring)	35.742	2, 698	< .001

Table 2 shows that scoring condition had a statistically significant effect on Economics essay test scores, $F(2, 698) = 35.742$, $p < .001$. This indicates that students' scores differed significantly depending on whether their responses were evaluated by human raters or Artificial Intelligence (AI)-assisted scoring. Relative to the AI-assisted scoring condition (reference category), both Human Rater 1 and Human Rater 2 produced significantly lower scores ($\beta = -0.452$, $p < .001$). The identical estimates for the two

human raters indicate that both human-rater conditions showed a similar scoring pattern compared with the AI-assisted scoring condition. Overall, the result suggests that scoring method significantly influenced the scores assigned to students' Economics essay responses, demonstrating systematic differences between human-rater and AI-assisted scoring conditions.

Table 3: Generalizability and Dependability Coefficients of Economics Essay Test Scores under Human-Rater and AI-Assisted Scoring Conditions

Reliability Metric	Value
Generalizability Coefficient ($E\rho^2$)	0.2773
Dependability Coefficient (Φ)	0.2773

The results presented in Table 3 show that the generalizability coefficient ($E\rho^2 = 0.2773$) and the dependability coefficient ($\Phi = 0.2773$) were both below the acceptable benchmark of 0.70. This indicates that the Economics essay test scores demonstrated low generalizability and dependability under the human-rater and Artificial Intelligence (AI)-assisted scoring conditions. The findings suggest that the current measurement design does not provide sufficient consistency and stability in estimating students' true performance. Therefore, the obtained scores may not be sufficiently reliable for making high-stakes decisions about students' Economics essay performance.

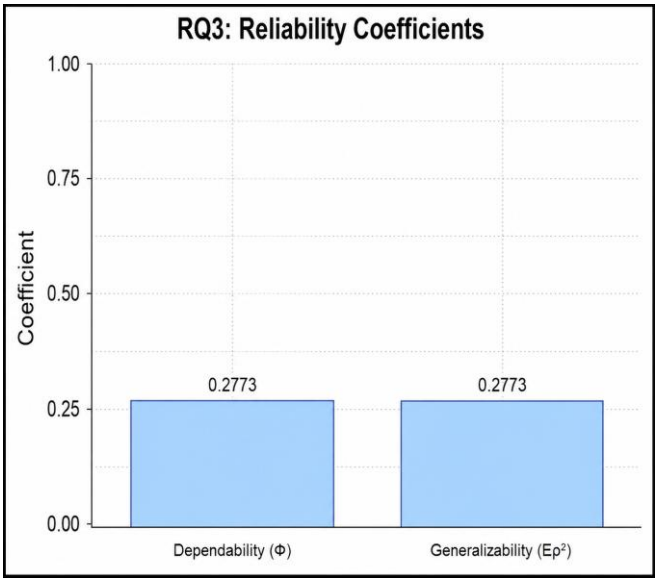

Figure 3

Table 4: Variance Components for Consistency Between Human-Rater and AI-Assisted Scoring of Economics Essay Responses

Component	Variance
ID × Item Interaction	0.4572
ID (Student)	0.0000
Item	0.0000
Residual	0.1130

Table 4 reveals that the largest source of variance was the student-by-item interaction effect (ID × Item = 0.4572), while the variance attributable to students alone (ID = 0.0000) and items alone (Item = 0.0000) was negligible. This indicates that variations in Economics essay test scores were mainly influenced by the interaction between individual students and specific essay items rather than stable differences among students or inherent differences among items. The residual variance (0.1130) represents unexplained variability remaining after accounting for the student-by-item interaction effect. Therefore, the consistency of human-rater and AI-assisted scoring appears to be influenced mainly by student–item response patterns.

The null hypothesis (H_{01}) was examined using the variance components and reliability coefficients obtained from the Generalizability Theory analysis. The results showed that error variance (0.48318) was the largest source of variation, followed by item variance (0.04044), while student variance (0.01340) was relatively small. In addition, the generalizability coefficient ($E\rho^2$ = 0.2773) and dependability coefficient (Φ = 0.2773) were both below the acceptable threshold of 0.70, indicating low reliability of the Economics essay test scores under the human-rater and Artificial Intelligence (AI)-assisted scoring conditions.

The null hypothesis (H_{02}) was tested using the linear mixed-effects model presented in Table 3. The results indicated that scoring condition had a statistically significant effect on Economics essay test scores, $F(2, 698) = 35.742$, $p < .001$. Relative to the AI-assisted scoring condition (reference category), both Human Rater 1 and Human Rater 2 produced significantly lower scores ($\beta = -0.452$, $p < .001$). Therefore, the null hypothesis was rejected, indicating a statistically significant difference between human-rater and Artificial Intelligence (AI)-assisted scoring conditions.

Discussion

The findings of Research Question One revealed that the largest variance component was attributable to measurement error, followed by item effects, while student variance was minimal. This indicates that score differences were influenced more by inconsistencies in the measurement process than by actual differences in students' abilities. The low generalizability coefficient ($E\rho^2$ = 0.2773) and dependability coefficient (Φ = 0.2773) further confirm the weak reliability and limited stability of the assessment scores. These findings are consistent with Jin et al. (2025), who reported that automated scoring performance is influenced by assessment characteristics, scoring criteria, and the alignment between scoring systems and intended measurement objectives. Similarly, Xue et al. (2025) observed that variability in automated scoring systems is affected by task characteristics, scoring prompts, and the representation of assessment criteria. However, the findings differ from Yeh and Lee (2024), who indicated that large language models can provide promising evidence of validity in automated essay scoring when appropriate validation procedures and scoring frameworks are applied. A possible explanation for these findings is that the inconsistency and limited discrimination power of the test items may have reduced the instrument's ability to distinguish among students of different ability levels. Variability in scoring interpretation, task ambiguity, and other uncontrolled factors in the scoring process may also have contributed to the dominance of measurement error and the instability of the scores.

The findings of Research Question Two revealed a statistically significant difference between human-rater and Artificial Intelligence (AI)-assisted scoring conditions, $F(2, 698) = 35.742$, $p < .001$, indicating that the scoring condition significantly influenced students' Economics essay test scores. Relative to the AI-assisted scoring condition, both Human Rater 1 and Human Rater 2 assigned significantly lower scores, suggesting systematic differences between human-rater and AI-assisted scoring. This finding supports Yamashita (2024), who reported that automated essay scoring systems may produce scoring patterns that differ from human raters because of differences in evaluation processes and scoring criteria. Similarly, Kundu and Barbosa (2024) found that large language models do not always produce essay evaluations identical to human judgments, particularly when responses involve complex reasoning and interpretation. However, this finding differs from Li and Ng (2024), who highlighted the increasing capability of automated essay scoring systems to achieve improved scoring performance when supported by appropriate computational methods and assessment frameworks. A possible explanation for this finding is the difference in scoring approaches, with human raters relying on contextual and qualitative judgment, whereas AI systems depend on learned language patterns and computational evaluation procedures. Variations in rubric interpretation, scoring

calibration, and response complexity may also have contributed to the observed differences in scores.

The findings of Research Question Three revealed that the generalizability coefficient ($E\rho^2$ = 0.2773) and dependability coefficient (Φ = 0.2773) were below the acceptable benchmark of 0.70, indicating low generalizability and dependability of the Economics essay test scores. This suggests that the instrument did not provide sufficient consistency in estimating students' true ability across measurement conditions. This finding agrees with Xue et al. (2025), who emphasized that the reliability of automated scoring depends on assessment design, scoring criteria, and the characteristics of constructed-response tasks. Similarly, Mughal et al. (2026) reported that automated scoring of constructed responses is influenced by item characteristics, response features, and scoring procedures, which may affect reliability outcomes. However, the finding differs from Yeh and Lee (2024), who suggested that large language models may provide useful scoring evidence when appropriate validation procedures are implemented. A possible explanation for this finding is that the limited number of test items and weak item discrimination reduced the instrument's ability to differentiate students' performance levels. In addition, the subjective nature of essay scoring and variation in students' responses may have increased measurement error, thereby reducing score dependability.

The findings of Research Question Four revealed that the student × item interaction accounted for the largest variance component, whereas the main effects of students and items were negligible. This suggests that students' performance depended largely on the interaction between individual students and specific essay items rather than on stable differences in students' ability or item characteristics. This finding agrees with Mughal et al. (2026), who demonstrated that constructed-response scoring performance is influenced by interactions among assessment tasks, response characteristics, and scoring procedures. Similarly, Xue et al. (2025) observed that differences in task characteristics and scoring conditions can influence variability in automated assessment outcomes. However, the finding differs from Yamashita (2024), who reported that automated scoring approaches can achieve acceptable consistency when scoring procedures are carefully structured. A possible explanation for this finding is that differences in item difficulty and ambiguity in scoring criteria may have influenced how students responded to individual items. In addition, the limited number of essay items and the complex nature of essay responses likely increased variability across student–item combinations, leading to the dominant interaction effect.

Conclusion

This study examined the sources of score variance, reliability, and consistency of Economics essay test scores under human-rater and Artificial Intelligence (AI)-assisted scoring conditions using Generalizability Theory and linear mixed-effects models. The findings revealed that measurement error constituted the largest source of variance, followed by item effects, while student variance was minimal. This indicates that the assessment instrument had limited ability to reliably differentiate students' true ability levels. Furthermore, the generalizability coefficient ($E\rho^2$ = 0.2773) and dependability coefficient (Φ = 0.2773) were below the acceptable benchmark of 0.70, indicating low generalizability and dependability of the scores. The results also revealed a statistically significant difference between human-rater and AI-assisted scoring conditions, suggesting that the scoring method influenced students' Economics essay test scores. In addition, the strong student × item interaction indicated that students' performance depended largely on specific student–item combinations rather than stable differences in students' ability or item characteristics. Overall, the findings suggest that improvements in test design, item quality, and scoring procedures are necessary to enhance the reliability and consistency of Economics essay test scores.

Recommendations

Based on the findings, the following recommendations are made:

1. Economics essay items should be improved to enhance clarity, discrimination, and reduce measurement error.
2. The number of well-constructed essay items should be increased to improve the generalizability and dependability of scores.
3. A hybrid approach combining human-rater and Artificial Intelligence (AI)-assisted scoring should be adopted to improve scoring consistency and reduce rater bias.
4. AI-assisted scoring systems should be properly trained and calibrated before use in high-stakes assessments.
5. Future studies should examine additional sources of score variation, such as rater training, scoring rubrics, and task characteristics, to improve the reliability of Economics essay test scores.

References

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

Engelhard, G. (2013). *Invariance of rater-mediated assessments: A measurement theory perspective*. Routledge.

Jin, H., Lima, C., & Wang, L. (2025). Automated scoring in learning progression-based assessment: A comparison of researcher and machine interpretations. *Educational Measurement: Issues and Practice, 44*(3), 25–37.

Kundu, A., & Barbosa, D. (2024). Are large language models good essay graders? *arXiv*. https://arxiv.org/abs/2409.13120

Li, S., & Ng, V. (2024). Automated essay scoring: Recent successes and future directions. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*.

Misgna, H., On, B. W., Lee, I., & Choi, G. S. (2025). A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review, 58*, Article 36.

Mughal, N., Imran, A. S., Daudpota, S. M., Kastrati, Z., & Noor, W. (2026). Exploring the potential of large language models for automated essay scoring in education. *Discover Artificial Intelligence, 6*, Article 166. https://doi.org/10.1007/s44163-026-01002-y

Williamson, D. M., Xi, X., & Breyer, F. J. (2020). A framework for automated scoring of constructed-response assessments. *Educational Measurement: Issues and Practice, 39*(1), 5–15.

Planning, Research and Statistics Department, Post Primary Schools Management Board, Nsukka Area Office. (2025). *Enrolment statistics of Senior Secondary School II Economics students in public secondary schools in Nsukka Education Zone for the 2025/2026 academic session* [Unpublished administrative record]. Post Primary Schools Management Board, Nsukka, Enugu State.

Xue, M., Liu, Y., Xiao, X., & Wilson, M. (2025). Automatic prompt engineering for automatic scoring. *Journal of Educational Measurement, 62*(4), 559–587.

Yamashita, T. (2024). An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. *Research Methods in Applied Linguistics, 3*(3), Article 100133.

Yeh, S., & Lee, C. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence, 6*, Article 100234.