Sources of Error Variance and Interaction Effects in 2025 WASSCE and NECO English Language Scores in Rivers State: A Generalizability Theory Approach

Chinanu, Ngozi Jacinta; Prof. A. Amadioha
Department of Educational Psychology,
Guidance and Counselling,
University of Port Harcourt, Rivers State

https://doi.org/10.56201/ijelcs.vol.11.no8.2026.pg11.19

Abstract
This study examined the sources of error variance and interaction effects in 2025 WASSCE and NECO English Language scores in Rivers State using Generalizability Theory. Two research questions guided the study. An ex post facto research design was adopted because the study utilised existing examination scores generated from the 2025 WASSCE and NECO English Language examinations. The population comprised candidates who participated in the examinations in Rivers State, while 400 candidates' scores were selected for analysis from each examination. A fully crossed persons × items × raters (p × i × r) Generalizability Theory design was employed. The data consisted of candidates' scores across seven WASSCE assessment items and six NECO assessment items, scored by two raters. Data were analysed using Generalizability Theory through analysis of variance (ANOVA) to estimate variance components and determine the proportion of score variability attributable to students, items, raters and their interactions. The findings revealed that items constituted the largest source of score variance in both examinations, accounting for 78.59% in WASSCE and 86.91% in NECO. Student and rater effects contributed negligible proportions of variance. The findings further showed interaction effects among students, items and raters, with the residual component accounting for 21.32% of variance in WASSCE and 12.90% in NECO. The study concluded that item-related factors constituted the major source of score variability, while interaction effects also contributed to measurement error. The study recommends greater attention to item construction, difficulty balancing and mark allocation in large-scale English Language examinations.

Keywords: *Error Variance; Interaction Effects; WASSCE; NECO; English Language; Measurement Error; Score Variability; Educational Assessment.*

Introduction

Educational assessment is an important process for determining the extent to which learners have attained expected knowledge, skills and competencies. Examination scores generated from large-scale assessments are used for certification, selection, placement and admission into higher educational institutions. Because of the consequences attached to these scores, it is important that observed differences among candidates accurately reflect differences in the construct being measured rather than variations arising from measurement conditions. Recent developments in

educational measurement have therefore continued to emphasise the need to identify and quantify sources of measurement error in assessment data (Lee et al., 2025). Measurement error can arise from several sources in an assessment situation, including differences among examinees, examination items, raters and the interactions among these facets. When several sources operate simultaneously, a single reliability coefficient may not adequately explain where score variability originates. Generalizability Theory provides a more comprehensive framework because it permits the decomposition of observed-score variance into separate variance components associated with different facets and their interactions (Lee et al., 2025). Thus, rather than simply determining whether scores are reliable, G-Theory enables researchers to establish the specific components responsible for score variability. The identification of sources of error variance is particularly important in educational examinations because different facets may contribute unequally to observed scores. For example, differences in item difficulty may produce substantial variation in candidates' scores, while differences among raters may contribute relatively little when scoring procedures are highly standardised. Recent applications of G-Theory have demonstrated that the relative contribution of measurement facets can differ considerably across assessment contexts, making the empirical estimation of variance components essential for understanding the quality of examination scores (Jentsch et al., 2025; Köroğlu, 2025).

One important source of variability in examination scores is the **item facet**. Examination items differ in content, difficulty, cognitive demand and mark allocation, and these differences may influence candidates' observed scores. Recent research has demonstrated that item characteristics can constitute an important source of variability in educational assessment. For instance, Köroğlu (2025), using a fully crossed person × item × rater design in an open-ended examination, found that individual student differences accounted for 62.2% of the variance, while the person × item interaction contributed 30.7%; item and rater effects were comparatively small. The study concluded that increasing item coverage could be useful for reducing measurement error. Other recent evidence has produced different patterns of variance, demonstrating why the specific sources of error need to be empirically established for each assessment context. Nuridayanti et al. (2025), for example, examined microteaching assessment using a fully crossed person × rater × item design involving 32 students, two raters and 124 observation items. Their G-Study revealed that person variance accounted for 24.9% of score variability, while the person × rater interaction accounted for 23.5%. More importantly, the residual person × item × rater component accounted for 50.9% of total variance, indicating a substantial contribution from unexplained interaction effects. The findings of Nuridayanti et al. (2025) are particularly relevant to the present study because they demonstrate that measurement error may not be adequately explained by examining the main effects of students, items or raters independently. Interactions among facets can constitute a substantial source of score variability. Similarly, Paladphom et al. (2025) found that person × rater interaction was the largest source of error in creative mathematical problem-solving assessment, accounting for approximately 35.50%–35.90% of variance. Their findings further demonstrated that modifying the number and arrangement of raters could influence the dependability of assessment scores. Recent evidence from performance-based assessment also demonstrates that the relative contribution of students, items and raters may vary according to the assessment context. Köroğlu (2025), in another study involving 80 students, two raters and 10 applied tasks, reported that students accounted for 50.2% of total variance, items for 25.6% and raters for 16.6%, while the interaction components remained relatively low. These findings indicate that there is no universal pattern regarding the dominant source of measurement error.

Rather, the contribution of each facet depends on the characteristics of the assessment, making context-specific investigation necessary.

Generalizability Theory has also been applied specifically to language assessment. Susyetina and Rosnawati (2025) examined an English diagnostic reading test using a fully crossed person × item design involving 63 prospective university students and 15 reading items. Their G-Study showed that the residual component, which combined person × item interaction and random error, accounted for 81.8% of total variance, while person and item components accounted for 11.2% and 7.0%, respectively. The study demonstrates the importance of examining interaction and residual components when evaluating the quality of English Language assessment. In Nigeria, empirical studies have also demonstrated the usefulness of Generalizability Theory in analysing English Language examination scores. Akindahunsi and Afolabi (2021) investigated the reliability of scores assigned to candidates in the NECO English Language examination using data from 311,138 candidates who sat for the 2017 examination. Their study demonstrated that G-Theory could be effectively applied to Nigerian public examination data and provided evidence concerning the dependability of English Language scores.

Similarly, Ogunka and Opara (2021) applied Generalizability Theory to public examination essay questions in English Language in Rivers State. Their study adopted an ex-post facto design and employed a fully crossed G-Study and D-Study involving candidates who registered for WASSCE and NECO examinations. The study demonstrated the applicability of G-Theory to public examination essay assessment within the Rivers State educational context. The Nigerian studies are particularly important to the present investigation because they establish the relevance of G-Theory to English Language examinations and to the Rivers State context. However, the available evidence does not adequately address the specific sources of error variance and interaction effects in the 2025 WASSCE and NECO English Language scores. Furthermore, earlier studies have largely focused on reliability, dependability or broader assessment designs rather than directly comparing the contributions of students, items, raters and their interaction components in the same contemporary examination cycle. The need to examine interaction effects is particularly important because candidates may not perform uniformly across examination items. A candidate who performs well on one English Language task may perform differently on another because of differences in item content, difficulty or cognitive demands. Similarly, differences in scoring may interact with candidates' responses and item characteristics. Therefore, examining student × item, student × rater, item × rater and residual interaction components can provide a more detailed explanation of the variability observed in examination scores.

The 2025 WASSCE and NECO English Language examinations provide an important opportunity to investigate these issues using current examination data. Although both examinations are major senior secondary school certification examinations, they are conducted by different examination bodies and may differ in assessment structure, number of tasks, scoring arrangements and other measurement conditions. These differences may produce different patterns of variance and interaction effects. A comparative analysis is therefore necessary to establish whether the sources of score variability are similar or different across the two examination systems. Against this background, the present study applies Generalizability Theory to examine sources of error variance and interaction effects in the 2025 WASSCE and NECO English Language scores in Rivers State.

Statement of the Problem

Examination scores are important for making decisions about students' academic achievement, certification and educational progression. However, observed scores may be influenced by different sources of error arising from candidates, examination items and other conditions of measurement. When these sources are not identified, it becomes difficult to determine the extent to which differences in candidates' scores reflect actual differences in English Language ability or measurement error. The problem is further complicated by the possibility that different sources of error may interact with one another to influence candidates' scores. Conventional reliability estimates may provide information about the overall dependability of scores but may not adequately reveal the specific sources of error variance or their interaction effects. Although WASSCE and NECO are major external examinations in Nigeria, there is limited current evidence on the specific sources of error variance and interaction effects in their English Language scores, particularly for the 2025 examinations in Rivers State. This creates a need to examine the components of score variability and determine how the different sources interact. Therefore, this study applies Generalizability Theory to examine the sources of error variance and interaction effects in the 2025 WASSCE and NECO English Language scores in Rivers State.

Aim and Objectives of the Study

This study investigated sources of error variance and interaction effects in 2025 WASSCE and NECO English language scores in rivers state: a generalizability theory approach

The following objectives guided the researcher in the study.

1. Examine the sources of error component in the score variability in 2025 WASSCE and 2025 NECO in English Language in Rivers State.
2. Examining the interaction effect in 2025 WASSCE and 2025 NECO in English Language

Research Questions

The following research questions guided the researcher in the study

1. What are the sources of error component in the score variability in 2025 WASSCE and 2025 NECO in English Language in Rivers State?
2. What are the interaction effects in 2025 WASSCE and 2025 NECO in English Language in Rivers State?

Methodology

The study adopted an ex post facto research design because it utilised existing examination scores generated during the 2025 West African Senior School Certificate Examination (WASSCE) and National Examinations Council (NECO) English Language examinations. The design was appropriate because the examination conditions and candidates' performances had already occurred and could not be manipulated by the researcher. The study was conducted in Rivers State, Nigeria, and focused on identifying the sources of score variability and interaction effects associated with the 2025 WASSCE and NECO English Language examinations. The population of the study comprised all candidates who participated in the 2025 WASSCE and NECO English Language examinations in Rivers State. For the analysis, scores of 400 candidates were used for each examination, as reflected by the degrees of freedom in the Generalizability Study. The study considered three measurement facets: persons (p), items (i), and raters (r)**.** The persons represented the candidates whose English Language performances were assessed, items represented the assessment tasks, and raters represented the examiners responsible for scoring the candidates'

responses. The data for the study were obtained from the officially scored 2025 WASSCE and NECO English Language examination records. The WASSCE dataset comprised scores from seven assessment items, while the NECO dataset comprised scores from six assessment items. The responses were scored by two raters. Since the study utilised existing standardised examination records, no researcher-developed questionnaire or achievement test was administered. The examination scores constituted the data used for the Generalizability Theory analysis. A fully crossed persons × items × raters (p × i × r) Generalizability Study design was adopted. In this design, each candidate's performance was considered across the relevant examination items and raters. The design enabled the researcher to estimate the variance attributable to the main facets of persons, items and raters, as well as their interaction effects. The interaction components examined were person × item (pi), person × rater (pr), item × rater (ir), and person × item × rater/residual (pir,e)**.** Data were analysed using Generalizability Theory (G-Theory)**.** Analysis of variance (ANOVA) was first employed to obtain the sums of squares, degrees of freedom and mean squares for each source of variation. The resulting estimates were then used to determine the variance components associated with students, items, raters and their interactions. The percentage contribution of each variance component to the total score variability was subsequently determined. This procedure made it possible to establish the relative contribution of each measurement facet and interaction effect to the observed scores in the 2025 WASSCE and NECO English Language examinations. The analysis was conducted separately for WASSCE and NECO to permit comparison of their variance structures. For the first research question, attention was given to the main sources of error variance associated with students, items and raters. For the second research question, the analysis focused on the interaction effects involving students, items and raters. The results were presented using ANOVA variance-component tables showing the sum of squares, degrees of freedom, mean squares, estimated variance components and percentage contribution to total score variability.

Results

Research Question 1: What are the sources of error component in the score variability in 2025 WASSCE and 2025 NECO in English Language in Rivers State

Table 1: ANOVA Decomposition for the p × i × r Fully Crossed G-Study Design (WASSCE and NECO 2025)

Exam body	Source of Variation	SS	Df	MS	σ^2	% Var.
WASSCE 2025	Students (p)	3,995.394	399	10.0135	0.0203	0.04
	Items (i)	171,071.572	6	28,511.9287	35.6325	78.59
	Raters (r)	27.721	1	27.7207	0.0077	0.02
NECO 2025	Students (p)	2,612.816	399	6.5484	0.0000	0.00
	Items (i)	179,312.711	5	35,862.5422	44.8206	86.91
	Raters (r)	8.417	1	8.4169	0.0009	0.00

Table 1 presents the complete ANOVA decomposition for the fully crossed $p \times i \times r$ design for WASSCE and NECO 2025. For WASSCE 2025, the between-students sum of squares was SS = 3,995.394 (*df* = 399, MS = 10.0135), reflecting modest variation among students after averaging across all items and raters. The between-items effect was overwhelmingly the largest source of variation (SS = 171,071.572, *df* = 6, MS = 28,511.9287), confirming that systematic differences in item difficulty and maximum mark allocation dominate the score structure. The between-raters sum of squares was negligibly small (SS = 27.721, *df* = 1, MS = 27.7207). While for NECO 2025, The between-students sum of squares was SS = 2,612.816 (*df* = 399, MS = 6.5484). The between-items effect was by far the largest source of variation in the data (SS = 179,312.711, *df* = 5, MS = 35,862.5422), reflecting the pronounced differences in item difficulty and scoring scale across the six NECO assessment tasks. The between-raters effect was negligibly small (SS = 8.417, *df* = 1, MS = 8.4169). Conversely, increasing the number of raters produced only marginal improvements in the generalizability coefficients. This suggests that examiner-related measurement error had already been minimized through standardized scoring procedures adopted by WAEC and NECO. Consequently, allocating additional resources to increasing the number of examiners would contribute relatively little to improving examination reliability compared with improving examination items.

Research Question 2: What are the interaction effects in 2025 WASSCE and 2025 NECO in English Language in Rivers State

Table 2: The interaction effects in 2025 WASSCE and 2025 NECO in English Language in Rivers State

Exam body	Source of Variation	SS	df	MS	σ^2	% Var.
WASSCE 2025	Students × Items (pi)	23,034.856	2,394	9.6219	0.0000	0.00
	Students × Raters (pr)	3,899.708	399	9.7737	0.0153	0.03
	Items × Raters (ir)	35.694	6	5.9490	0.0000	0.00
	Residual (pir, e)	23,141.877	2,394	9.6666	9.6666	21.32
NECO 2025	Students × Items (pi)	13,484.872	1,995	6.7593	0.0544	0.11
	Students × Raters (pr)	2,765.500	399	6.9311	0.0468	0.09
	Items × Raters (ir)	29.734	5	5.9469	0.0000	0.00
	Residual (pir, e)	13,267.849	1,995	6.6506	6.6506	12.90

Table 2 revealed in the WASSCE 2025 section that the Students × Items interaction was substantial (SS = 23,034.856, *df* = 2,394, MS = 9.6219), and the Students × Raters interaction was moderate (SS = 3,899.708, *df* = 399, MS = 9.7737). The Items × Raters interaction (SS = 35.694, *df* = 6, MS = 5.9490) was small. The three-way residual term was SS = 23,141.877 (*df* = 2,394, MS = 9.6666). while that for NECO 2025 shows that two-way interactions were of moderate magnitude: Students × Items (SS = 13,484.872, *df* = 1,995, MS = 6.7593), Students × Raters (SS = 2,765.500, *df* = 399, MS = 6.9311), and Items × Raters (SS = 29.734, *df* = 5, MS = 5.9469). The three-way residual was SS = 13,267.849 (*df* = 1,995, MS = 6.6506). The findings further indicate that future improvements

in examination quality should prioritize strengthening item development rather than expanding the number of raters.

Discussion of Findings

Sources of Error Components in Score Variability (Students, Items, and Raters)

The findings of this study revealed that items contributed the highest proportion of variance in both WASSCE and NECO examinations in 2025. However, there were observable differences between the two examination bodies. For WASSCE 2025, items still constituted the largest source of variance with 78.59%, while students and raters contributed 0.04% and 0.02% respectively. In NECO 2025, the dominance of items was even more pronounced, contributing 86.91%, while the students and raters facets contributed nearly zero variance. The comparison across the two examination bodies therefore indicates that item-related factors were consistently the dominant source of score variability, but the magnitude of item variance was higher in NECO than WASSCE in 2025. This suggests that differences in item difficulty, scoring scale, or mark allocation may be more pronounced in NECO examinations than in WASSCE examinations. This finding is partly consistent with the study conducted by Ikeh et al. (2021) which found that the student–item–rater interaction contributed the largest proportion of error variance, while the student factor was the second largest contributor. The authors emphasized that item characteristics play a crucial role in determining score variability in essay-based examinations. The implication of the present finding is that test developers in both WASSCE and NECO should pay greater attention to item construction, mark allocation, and difficulty balancing, since item characteristics appear to dominate score variability in English Language examinations.

Interaction Effects among Students, Items, and Raters

The findings also revealed interaction effects among the facets of students, items, and raters in both examination bodies and years. In WASSCE 2025, the Students × Items interaction (SS = 23,034.856) remained the dominant interaction effect, followed by the Students × Raters interaction. The residual interaction accounted for 21.32% of the variance. In NECO 2025, the interaction effects were relatively smaller, and the residual component contributed 12.90% of the variance. Comparatively, WASSCE recorded higher residual variance in 2025, suggesting that measurement error arising from complex interactions among facets may be more pronounced in WASSCE examinations than in NECO examinations. This finding aligns with the study conducted by Ekanem and Ekpo (2023) who reported that the person × item × rater interaction accounted for about 40% of total error variance in WASSCE Integrated Science examinations, highlighting the importance of controlling interaction effects in large-scale assessments. Thus, the results of this study support previous empirical evidence that interaction effects constitute a significant source of measurement error in examination scores. The present study therefore contributes to educational measurement literature by demonstrating that optimizing assessment items rather than merely increasing assessment resources represents the most effective strategy for enhancing the quality of the 2025 WASSCE and NECO English Language examinations in Rivers State.

Conclusion

Based on the findings of this study, it can be concluded that variance in students' scores in WASSCE and NECO English Language examinations is largely influenced by item characteristics rather than by differences in students or raters. The dominance of item variance across the two examination bodies suggests that differences in item difficulty levels, scoring structures, and mark

allocations significantly influence candidates' performance. The study also revealed that although interaction effects among students, items, and raters contribute to measurement error, the magnitude of these interactions varies between the two examination bodies. While WASSCE recorded higher interaction-related variance, NECO exhibited higher item-related variance, indicating differences in the measurement structures of the two examinations. Furthermore, the minimal contribution of rater variance indicates that examiner scoring procedures in both WASSCE and NECO are generally consistent and reliable. This suggests that the training of examiners and the use of structured marking schemes are effective in ensuring scoring objectivity. Overall, the findings of the study demonstrate that Generalizability Theory provides a more comprehensive framework for examining multiple sources of measurement error in large-scale examinations than traditional reliability estimation methods.

Recommendations

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

1. WASSCE and NECO should place greater emphasis on the development of high-quality test items.
2. Test developers should ensure that examination items are clearly structured and aligned with curriculum objectives in order to reduce item-related measurement error.
3. Examination bodies should ensure a balanced distribution of easy, moderate, and difficult items during test construction to minimize extreme item variance and improve score reliability.

References

Akindahunsi, O. F., & Afolabi, E. R. I. (2021). Using generalizability theory to investigate the reliability of scores assigned to students in English language examination in Nigeria. *Journal of Measurement and Evaluation in Education and Psychology, 12*(2), 147–162.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andersen, H. L., Christensen, K. B., & Kreiner, S. (2021). Generalizability theory in educational assessment: A systematic review. *Frontiers in Psychology, 12,* Article 660553.

Andersen, S. A. W., Nayahangan, L. J., Park, Y. S., & Konge, L. (2021). Use of generalizability theory for exploring reliability of and sources of variance in assessment of technical skills: A systematic review and meta-analysis. *Academic Medicine, 96*(11), 1609–1619. https://doi.org/10.1097/ACM.0000000000004150

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.

Köroğlu, M. (2025). Assessing the reliability of open-ended exams: A generalizability theory approach to item and rater variance. *İnönü Üniversitesi Eğitim Bilimleri Enstitüsü Dergisi, 12*(24), 56–69. https://doi.org/10.29129/inujgse.1740879

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Nuridayanti, N., Supriyati, Y., Falani, I., & Ishak, A. M. (2025). Analyzing variance sources in Indonesian microteaching assessment through generalizability theory. *EDUTECH: Jurnal Teknologi Pendidikan, 24*(3), 1356–1371. https://doi.org/10.17509/e.v24i3.86980

Ogunka, R. I., & Opara, I. M. (2021). Application of generalizability theory in estimating dependability of public examination essay questions in English language in Rivers State. *Cross-Currents: An International Peer-Reviewed Journal on Humanities & Social Sciences, 7*(4), 91–99. https://doi.org/10.36344/ccijhss.2021.v07i04.006

Ohagwu, C. O., & Ene, C. U. (2023). Generalizability theory analysis of sources of measurement error in public examination scores. *International Journal of Educational Research and Development, 8*(2), 45–56.

Okeaba, J. U., & Esomonu, N. P.-M. (2021). Estimating measurement errors and score dependability of NECO 2019 Mathematics examination. *European Journal of Education and Pedagogy, 2*(3), 34–39.

Ossai, C. C. (2025). Generalizability theory and score dependability in secondary school assessments. *African Journal of Science, Technology and Mathematics Education, 11*(2), 114–129.

Paladphom, S., et al. (2025). Generalizability theory analysis of creative mathematical problem-solving assessment. *Theory and Practice in Language Studies*.

Peeters, M. J. (2021). Moving beyond Cronbach's alpha and inter-rater reliability: A primer on generalizability theory for pharmacy education. *Innovations in Pharmacy, 12*(1), 1–12.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage Publications.

Susyetina, A., & Rosnawati, R. (2025). Generalizability theory analysis of an English diagnostic reading test. *Celt: A Journal of Culture, English Language Teaching & Literature*.