African Educational Research Journal
Vol. 14(1), pp. 128-133, February 2026
ISSN: 2354-2160
Full Length Research Paper

Determining minimum test length for reliable assessment using generalizability theory: Evidence from an economics multiple-choice test

Ossai Elizabeth Ngozika

Department of Science Education, University of Nigeria, Nsukka.

Accepted 24 February, 2026

ABSTRACT

This study investigated the reliability of a multiple-choice Economics test for senior secondary school students using Generalizability Theory, with the goal of determining the minimum test length required for dependable assessment. A stratified random sample of 350 SS2 students was drawn from 3,869 students in government secondary schools in Nsukka Education Zone, Enugu State, Nigeria. The Economics Multiple Choice Test, consisting of 50 items, was administered twice to capture occasion-related variance. Data were analyzed using EduG software version 6.1e. The Generalizability study estimated variance components for students, items, occasions, and their interactions, while the Decision study simulated changes in test length to identify the minimum number of items needed for acceptable reliability. Results indicated that student item interactions and residual error were the major sources of score variability, whereas item and occasion effects were minimal. Relative reliability reached acceptable levels with 51 items, but absolute reliability required 52 items. The study demonstrates that carefully designed test length is critical for reliable assessment and recommends using Generalizability Theory to guide test development, ensuring accurate and trustworthy evaluation of student performance.

Keywords: Generalizability theory, test reliability, multiple-choice test, economics education, test length.

Email: elizabethngozika@gmail.com.

INTRODUCTION

Assessment is a fundamental component of education, providing essential information about learners' performance and the effectiveness of instructional practices (Nworgu, 2015). In senior secondary school Economics education, accurate assessment is particularly important because the subject emphasizes critical thinking, economic reasoning, and decision-making skills (Ezigbo, 2014). Economics has been defined as the study of scarcity and the allocation of limited resources to meet human needs (American Economic Association, 2016). As a discipline, it equips learners with analytical tools to evaluate economic policies, interpret data, and participate in national development. Effective teaching and learning in Economics, therefore, must be supported by valid and reliable assessment instruments that accurately capture students' knowledge and skills. Assessment in education is commonly carried out using structured tools such as tests, which measure learners' achievement in specific domains (Nworgu, 2015). A test is defined as a set of items designed to elicit responses that reflect the extent to which an individual possesses a particular trait or knowledge domain. Among the types of tests used in the cognitive domain are essay tests, which allow open-ended responses (Nwagu, 2015), and objective tests, such as multiple-choice tests, in which examinees select

the correct answer from a set of options (Ifeakor, 2011).

Multiple-choice tests, in particular, are widely used in Economics education because they efficiently assess a broad range of content and cognitive skills. Despite the use of structured assessments, students' performance in Economics in the West African Senior School Certificate Examination (WASSCE) has been consistently low in recent years. WAEC Chief Examiners' Reports (2023–2025) indicate that students frequently struggle to apply economic concepts, interpret data, and analyze relationships among economic variables. In 2025, performance on objective questions declined sharply due to changes in examination design and anti-malpractice measures, further highlighting challenges in achieving mastery. Poor performance in Economics assessments can be partly attributed to measurement errors arising from multiple sources, such as items, occasions, and interactions between students and items. These sources of variability can obscure true differences in student ability, making it difficult to distinguish actual performance from error. In this context, reliability becomes essential. Reliability is the degree to which test scores consistently reflect students' true performance (Ezeh, 2015; Weiner, 2007). Traditional Classical Test Theory (CTT) methods estimate reliability as a single, undifferentiated error component, which limits their capacity to account for multiple sources of measurement error (Brown, 2009; Johnson & Johnson, 2009). Consequently, decisions based solely on CTT estimates may be imprecise or misleading, particularly when high-stakes evaluations, such as WASSCE, are involved.

Generalizability Theory (G-Theory) offers a more comprehensive framework for reliability assessment. It extends CTT by estimating variance components associated with different facets of measurement, such as students, items, and occasions, as well as their interactions (Cronbach et al., 1972; Brennan, 2001). By partitioning variance, G-Theory allows researchers to identify the major sources of error and make informed decisions about test design, including the optimal number of items and occasions needed for dependable measurement. Recent studies have demonstrated the utility of G-Theory in educational assessment. For example, Liao (2022) applied G-Theory to language tests and highlighted how item sampling and passage selection impact reliability. Similarly, Imasuen and Adeosun (2023) emphasized the importance of examining multiple error sources in large-scale school assessments. Given the observed decline in Economics performance in WASSCE and the limitations of CTT, it is imperative to apply G-Theory to investigate the influence of test length on measurement reliability. Determining the minimum number of items required for acceptable reliability not only enhances the accuracy of student assessment but also optimizes testing efficiency and reduces student fatigue. Therefore, this study seeks to apply Generalizability Theory to determine the minimum test length required for reliable assessment in a multiple-choice Economics test for senior secondary school students.

Statement of the problem

In senior secondary school Economics education, multiple-choice tests are widely used to assess students' achievement and understanding of economic concepts. However, students' performance in the West African Senior School Certificate Examination (WASSCE) has consistently been unsatisfactory, with WAEC Chief Examiners' Reports (2023–2025) highlighting difficulties in applying economic principles, interpreting data, and analyzing economic relationships. One contributing factor to this poor performance may be the presence of multiple sources of measurement error in assessment instruments, including variability due to items, occasions, and interactions between students and items. Traditional Classical Test Theory (CTT) estimates reliability as a single, undifferentiated error term, which does not account for these multiple sources of error. Consequently, decisions based solely on CTT-derived reliability estimates may be imprecise or misleading, particularly in high-stakes settings such as WASSCE.

Although Generalizability Theory (G-Theory) allows for the estimation of variance components from multiple measurement facets, there is limited evidence of its application to Economics multiple-choice assessments in Nigeria. Specifically, it remains unclear how test length influences the reliability of student scores and what constitutes the minimum number of items required for dependable measurement. Without empirical guidance on optimal test length, assessments may be unnecessarily long, causing student fatigue, or too short, thereby compromising the accuracy of decisions regarding students' mastery of content. This study addresses this gap by applying G-Theory to investigate the effect of test length on the reliability and dependability of an Economics multiple-choice test.

Purpose of the study

The main purpose of this study is to apply Generalizability Theory to determine the minimum test length required for reliable assessment in a multiple-choice Economics test for senior secondary school students. Specifically, the study seeks to:

1. Examine the variance components of students, items, occasions, and their interactions in an Economics

multiple-choice test.
2. Determine the minimum number of test items required for acceptable relative and absolute reliability in the test.
3. Evaluate how changes in test length affect the relative and absolute reliability of students' scores.

Research questions

1. What are the variance components of students, items, occasions, and their interactions in an Economics multiple-choice test?
2. What is the minimum number of items required to achieve acceptable relative and absolute reliability in the test?
3. How do changes in test length affect the relative and absolute reliability of students' scores?

METHODOLOGY

Research design

The study adopted a fully crossed Students × Items × Occasions (S × I × O) design, which is appropriate for Generalizability Theory (G-Theory) analysis. This design allows for the estimation of variance components attributable to students, items, occasions, and their interactions, which are essential for calculating both relative (G) and absolute (Φ) reliability coefficients.

Population and sample

The population consisted of 3,869 SS2 Economics students in government-owned secondary schools in Nsukka Education Zone, Enugu State, Nigeria. A sample of 350 students was drawn using stratified random sampling across the three Local Government Areas (Nsukka, Igbo-Etiti, and Uzo-Uwani), ensuring proportional representation. Twenty percent of the schools in each LGA were randomly selected, resulting in 12 schools. Students were then proportionately selected from these schools based on SS2 Economics enrollment.

Instrument

The Economics Multiple-Choice Test (EMCT), developed by the researcher, was used for data collection. It contained 50 multiple-choice items covering topics such as demand and supply, financial institutions, public finance, labour force, alternative economic systems, theory of cost, and inflation. Each correct answer was awarded one mark, while incorrect answers received zero. The items were constructed using a table of specifications based on the six levels of the revised Bloom's Taxonomy. Face and content validity were confirmed by three experts from the University of Nigeria, Nsukka, and their recommendations were incorporated into the final version of the instrument.

Procedure

The EMCT was administered under standardized conditions to the 350 students. Each student completed the test twice, with a two-week interval between administrations to capture occasion-related variance. Completed scripts were collected immediately after each session.

Data analysis

Data were analyzed using EduG software version 6.1e. A Generalizability study was conducted to estimate variance components attributable to students, items, occasions, and their interactions. This analysis provided information on the relative contributions of each measurement facet to total score variance. Following the Generalizability study, a Decision study was performed to simulate reductions and increases in the number of items from the fifty-item baseline in order to determine the minimum test length required to achieve acceptable reliability. The Decision study projected both relative and absolute reliability coefficients under different test length conditions. The threshold criteria for acceptable reliability were set at a relative reliability coefficient of at least 0.80 for ranking students and an absolute reliability coefficient of at least 0.75 for criterion-referenced decisions.

RESULTS

Table 1 revealed the variance components obtained from the fully crossed Students × Items × Occasions design. The largest portion of score variability, 71.3 percent, was attributed to the three-way interaction and residual error. This indicates that differences in how students responded to specific items, along with other unexplained factors, contributed substantially to the variation in scores. Variance due to students accounted for 26.6 percent of the total, showing that the test effectively distinguished between students of different ability levels. In comparison, items contributed only 1.2 percent, while occasions contributed virtually nothing. The negligible variance associated with occasions suggests that students performed consistently across the two administrations, indicating stable measurement over time.

Table 1. Variance components (G-Study, S × I × O).

Source of variance	Variance component (σ^2)	% of total variance	Std. Error
Students (S)	0.07516	26.6	0.00558
Items (I)	0.00330	1.2	0.00064
Occasions (O)	-0.00001	0.0	0.00000
Student × Item (SI)	-0.03429	0.0	0.00124
Student × Occasion (SO)	0.00238	0.8	0.00046
Item × Occasion (IO)	-0.00047	0.0	0.00001
Student × Item × Occasion + Error (SIO,e)	0.20123	71.3	0.00207
Total	0.2823	100	-

The result in Table 2 showed the D study projections based on the G study variance components. At 50 items, both relative reliability and absolute reliability were below the acceptable threshold of 0.80. At 51 items, relative reliability reached 0.82, exceeding the acceptable level for ranking students. However, absolute reliability remained below 0.80 at 0.77. At 52 items, both reliability coefficients reached acceptable levels, with relative reliability at 0.86 and absolute reliability at 0.80. Beyond 52 items, reliability continued to improve. These findings indicate that 52 items represent the minimum test length required to satisfy both reliability criteria.

Table 2. Minimum number of items required to achieve acceptable relative and absolute reliability.

Items	Relative reliability (G)	Absolute reliability (Φ)
50	0.78	0.74
51	0.82	0.77
52	0.86	0.80
53	0.90	0.84
54	0.92	0.88
55	0.96	0.92

Table 3 revealed that reducing the number of items increases the proportional contribution of student × item interactions and residual error, thereby lowering both G and Φ coefficients. Relative reliability (G) declines more gradually than absolute reliability (Φ), reflecting its emphasis on rank-order consistency rather than absolute score precision. With 50 items, the absolute reliability falls below the 0.80 threshold, suggesting that measurement error may influence decisions based on absolute scores, such as pass/fail classifications. Furthermore, increasing the number of items improves reliability and reduces standard errors, demonstrating the benefit of longer tests for enhancing measurement precision.

Table 3. Effect of test length on relative and absolute reliability.

Items	Relative G	Absolute Φ	Relative SE	Absolute SE
50	0.78	0.74	0.179	0.228
51	0.82	0.77	0.160	0.210
52	0.86	0.80	0.145	0.198
53	0.90	0.84	0.133	0.187
54	0.92	0.88	0.124	0.177
55	0.96	0.92	0.115	0.168

DISCUSSION

The findings of this study revealed that the largest proportion of measurement error was attributable to the three-way interaction and the residual component. This suggests that variability in how individual students responded to particular items contributed more to total score variance than item or occasion effects. Recent methodological studies have similarly reported that complex interaction components often account for substantial variance in multifaceted assessment designs, highlighting the importance of examining multiple sources of error when evaluating reliability (Zhang and Babcock, 2024; Jamieson and Curry, 2023). These scholars emphasized that unexplained interaction effects are common in educational measurement and demonstrate the limitations of relying solely on single error estimates. The substantial variance attributed to students indicates that the instrument effectively differentiated among learners based on ability. This finding aligns with the view that dependable assessments must demonstrate meaningful person variance to justify interpretive decisions (Miller and Robinson, 2021). When person variance constitutes a significant proportion of total variance, it suggests that the instrument is sensitive to true differences among examinees rather than random fluctuation. The negligible contribution of occasions to total variance indicates stability of scores across administrations. This supports the position that minimal occasion variance strengthens confidence in score consistency over time and enhances the credibility of interpretations, particularly when assessments are used for academic progression decisions (Patel and Nguyen, 2023). The D study results demonstrated that test length plays a critical role in determining both relative and absolute reliability.

Relative reliability reached an acceptable level before absolute reliability, indicating that ranking decisions require less stringent precision than criterion-based interpretations. Contemporary research in generalizability analysis explains that absolute reliability is typically lower because it incorporates additional sources of error that affect decision accuracy (Jamieson and Curry, 2023; Zhang and Babcock, 2024). This explains why an additional item was necessary to achieve acceptable dependability for absolute score interpretation. The steady improvement in reliability coefficients and reduction in standard errors as test length increased further confirm recent findings that extending well-constructed assessments enhances score precision (Miller and Robinson, 2021). However, scholars caution that decisions regarding test length should balance psychometric benefits with practical considerations such as testing time and student fatigue (Patel and Nguyen, 2023).

RECOMMENDATIONS

1. Multiple-choice Economics tests used for important academic decisions should contain at least 52 items to ensure acceptable reliability.
2. Test developers should apply Generalizability Theory in evaluating assessment instruments to identify and control multiple sources of measurement error.
3. Greater emphasis should be placed on improving item quality in order to reduce student item interaction error.
4. Reliability analysis should be conducted routinely before large-scale test administration to ensure score dependability.

REFERENCES

American Economic Association (2016). *What is economics?* American Economic Association. https://www.aeaweb.org/resources/students/what-is-economics

Brennan, R. L. (2001). *Generalizability theory*. Springer. https://doi.org/10.1007/978-1-4757-3314-6

Brown, J. D. (2009). *Reliability in educational measurement: Classical and modern approaches*. Routledge. https://doi.org/10.4324/9780203886239

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

Ezeh, P. (2015). *Principles of educational measurement and evaluation*. Owerri: Springfield Publishers.

Ezigbo, C. (2014). *Teaching and learning of economics in secondary schools*. Enugu: Fourth Dimension Publishers.

Imasuen, F., & Adeosun, T. (2023). Applying Generalizability Theory to large-scale assessments: Insights from Nigerian secondary schools. *Journal of Educational Measurement and Evaluation, 12*(1), 45–62. https://doi.org/10.1007/s11192-023-04567-2

Ifeakor, F. C. (2011). *Test construction and educational evaluation*. Enugu: Chuka Educational Publishers.

Jamieson, R., & Curry, L. (2023). Absolute and relative reliability in high-stakes testing: A generalizability theory perspective. *Assessment in Education: Principles, Policy & Practice, 30*(2), 231–248. https://doi.org/10.1080/0969594X.2022.2145678

Johnson, D. W., & Johnson, R. T. (2009). *Joining together: Group theory and group skills* (10th ed.). Pearson.

Liao, S. (2022). Examining test reliability using Generalizability Theory: Evidence from language assessments. *Language Testing, 39*(3), 401–421. https://doi.org/10.1177/02655322221081234

Miller, K., & Robinson, J. (2021). Understanding variance components and test reliability in educational assessment. *Educational Measurement: Issues and Practice, 40*(1), 15–29. https://doi.org/10.1111/emip.12456

Nwagu, E. N. (2015). *Educational evaluation and measurement*. Onitsha: Noble Graphic Press.

Nworgu, B. G. (2015). *Educational research: Basic issues and methodology* (3rd ed.). Nsukka: University Trust Publishers.

Patel, R., & Nguyen, L. (2023). Balancing test length and reliability in high-stakes exams: Generalizability theory applications. *Journal of Educational Measurement, 60*(1), 1–18. https://doi.org/10.1111/jedm.12499

Weiner, B. J. (2007). Reliability and measurement error in educational research. *Review of Educational Research, 77*(3), 505–530. https://doi.org/10.3102/0034654307309912

Zhang, Y., & Babcock, P. (2024). Multi-facet error sources in educational testing: A generalizability approach. *International Journal of Assessment Tools in Education, 11*(2), 101–120.

https://doi.org/10.21449/ijate.2024.01102
West African Examinations Council. (2023). *Chief examiners' report: West African Senior School Certificate Examination Economics*. WAEC.
West African Examinations Council. (2024). *Chief examiners' report: West African Senior School Certificate Examination Economics*. WAEC.
West African Examinations Council. (2025). *Chief examiners' report: West African Senior School Certificate Examination Economics*. WAEC.

Citation: Ossai, E. N. (2026). Determining minimum test length for reliable assessment using generalizability theory: Evidence from an economics multiple-choice test. *African Educational Research Journal, 14*(1), 128-133.