ESTIMATING WEST AFRICAN EXAMINATION COUNCIL MATHEMATICS THEORY TEST ITEMS OVER MULTIPLE FACETS USING GENERALIZABILITY THEORY

Catherine U Ene[1], Sylvanus C. Ohagwu[1], Basil C. Oguguo[1], Mercy Ngozi Nwoye[1], Francis Elochukwu Ikeh[1], Felicia Chinyere Ugwu[1], Janehilda Oluchi Agugoesi[1]

[1]Department of Science Education, University of Nigeria, Nsukka, Enugu State, Nigeria

Correspondence: Francis Elochukwu Ikeh, Department of Science Education, University of Nigeria, Nsukka, Enugu State, Nigeria (elochukwu.ikeh@unn.edu.ng)

Abstract

The study investigated the application of generalizability theory in estimating West Africa Examination Council (WAEC) mathematics test scores over multiple facets. The study employed random effects, two-facets fully crossed design for a G-study and D-study. A sample of 898 senior secondary three (SS3) students was used for the study. Four research questions guided the study. Mathematics Achievement Essay Test (MAET) was used to collect data for the study.The instrument was validated by three experts in Science Education Department (one in Mathematics Education Unit and two in Measurement and Evaluation Unit), Faculty of Education, University of Nigeria Nsukka, Enugu State, Nigeria. The reliability of the instrument was established using Kendal coefficient of concordance (W) and reliability coefficient of 0.91 was obtained. A computer program EduG version 6.1-e based on the Analysis of Variance (ANOVA) and Generalizability Theory (GT) was used to carry out the Generalizability analysis. The resultrevealed that largest contribution of error variance was from the students followed by student-by-question-by-rater factor. The third largest variance component was student-by-question interaction followed by variance of the rater. The variances due to question, student-by-rater and question-by-rater interaction were zero each.

***Keywords**: Generalizability Theory, Mathematics, West African Examination Council*

Introduction

Mathematics is one of the required courses in Nigerian primary and secondary schools. Mathematics is one of the fields of science that helps to progress current science and technology around the world. Even among mathematicians, mathematics is one of the oldest sciences for which there is no universally accepted definition.Many authors have defined mathematics from various angles. Yadav(2017), for example, defines mathematics as the study of assumptions, their properties, and applications. Mathematics, according to Elaine (2013), is a science that works with logic, shapes, amount, and arranging. Mathematics aids in the development of reasoning skills in students. The majority of scientific and commercial research and development is based on mathematics (Kusmaryono, 2014). As a result, Enu, Osei, and Nkum(2015) described mathematics as the foundation of scientific and technological knowledge that is critical to a country's social-economic growth.According to Asikhia (2013), mathematics is a large topic that deals with the measurement, qualities, and relationships of quantities expressed in numbers or symbols.

The importance of mathematics cannot be overemphasized. Mathematics has a significant impact on how people interact with their private, social, and civic lives (Anthony &Walshaw, 2009). This may be the reason why Eraikhuemen (2003) posits that a disciplined and orderly pattern of life can only be accomplished through the culture of mathematics because of the vast uses of mathematics.According to Valt and Maree (2007), mathematics is an important topic in secondary school and an adequate learning facility that is critical in any country. Therefore, Mathematics is a crucial tool in the advancement of science and technology. Despite the importance of Mathematics as a subject, over the years, students' performance in this vital subject has not been encouraging.This is evidence from WAEC Chief Examiners' Reports of 2014-2018. For instance, from the report, students recorded an unencouraging percentage pass of 31.28%, 36.68%, 52.97%, 50.22% and 49.98% respectivelyfor 2014-2018.

It is an undeniable reality that the success of learning a subject is determined by a variety of factors. Such factors include; poor primary school background (Obioma, 2006), lack of instructional resources (Yara & Otieno, 2010), poor teaching and learning environment (Abakpa &Iji, 2011), and inadequate method of instruction (Kalijah, 2002, Agomuo&Nzewi, 2003).Among several factors affecting the performance of test takers in a test is the sequence of the test items (Soureshjani, 2011). This implies that test items are also one of the most controversial issues which influence performance of students in an examination.In support of the above statement, Jumaeda (2017) opines that reasonable number of students encountered difficulties in responding to test items as a result of the structure of the test among others. Moreover, rater variability has been identified as one of the most beneficial and impediment factors which influence students test scores especially essay test.

It's no surprise that, according to Adeyemi (2010), teachers have a crucial influence in determining scores that indicate students' academic progress. In the same vein, Uzun, Aktaş, Aşiret and Yorulmaz (2018) opined that raters are also a source of variability in performance ratings. Raters play a crucial role in performance evaluations. This is due to the fact that raters' techniques, methods and approaches differ, affecting their ratings. Raters may deliver unexpected scores given genuine examinee competence due to a variety of variables such as insufficient training, misalignment of scoring with the rubric's worldview, or distractions while scoring (Myford& Wolfe, 2002; 2004). The degree of possible rater effect influences the appropriateness of using evaluative rubrics.

Moreover, students being an object of measurement are seen as a source of variance through which every other facet has imparted on. Individual sources of variability are individuals' capacity to demonstrate talent or knowledge in relation to a problem, task, or product (Uzun, Aktaş, Aşiret&Yorulmaz, 2018). It is assumed that the talents of those being assessed will differ from one another. Students, on the other hand, are one of the most important determinants of mathematics test scores in Nigeria and around the world. According toMohamed andWaheed(2011), one crucial component that has been regularly researched is students' attitude toward mathematics. The attitude of students toward mathematics is critical in the teaching and learning of mathematics.

In an attempt to identify these facets (items, raters, and persons) in a single analysis, there is need to employ Generalizability theory (G theory). This is because. G theory expands on classical test theory by re-conceptualizing the undifferentiated error component into different sources of systematic variability and random errors (Brennan, as cited in Uzun, Aktaş, Aşiret&Yorulmaz, 2018).One of the advantages of generalizability theory over classical test theory in performance evaluation is the capacity to analyze errors from multiple sources at the same time (Brennan, 2001). In line with the statement, Shavelson and Webb as cited in Semmelroth (2013) opined that the advantage of G theory is that it allows numerous sources of measurement error to be assessed separately in a single analysis. As a result of generalization theory, a more realistic error determination may be established, and more realistic conclusions and conclusions can be reached.According to Li, Shavelson, Yin and Wiley(2015), generalizability theory is a psychometric theory that relies on a statistical sampling strategy to divide scores into their respective score variations. Generalizability (G) theory partitions scores into their underlying multiple sources of variation.

GT equally selects the most appropriate measuring settings in order to produce reliable results. GT distinguishes between generalizability (G) and decision (D) research in this regard. Researchers in G studies first identify measurement facet(s) that may be of interest and/or have a significant impact on observed measurements (Shavelson & Webb, 2005). After that, the variance components of the main and interaction effects relevant to the measurement items and facet(s) are estimated.In D studies, G-study results are to be generalized from a particular measurement procedure, usually the data collection procedure at hand, to other measurement procedures (Lin, 2014). Specifically, D studies provide information as to how score reliability changes when the number of ratings/raters as well as items increases or decreases.As a result, generalizability theory, which liberalizes classical test theory, is required in order for a researcher to quantify and distinguish the sources of discrepancies in observed scores that emerge, or could emerge, in any measurement process. These observations underscore the need to apply generalizability theory to estimateWest Africa Examination Council Mathematics test items over multiple facets.

Methodology

The study adopted a random effect, two-facets fully crossed S×Q×R G and D study designs. Fully crossed design was used in order to estimate all the possible variance components in the measurement situation. The population of the study was 9040 subjects (64 Mathematics teachers and 8976 senior secondary three (SS 3) Mathematics students) in the 54 senior secondary schools in Udi education zone of Enugu State. The sample size of 898 SS3 students and 6 mathematics teachers was used for the study. In determining the sample size of the study, simple random sampling technique was used to sample three schools in each of the two Local government areas that constitute Udi education zone of Enugu state. The number of SS3 Mathematics students and teachers in the six sampled schools was used for the study.

The instrument used for data collection was Mathematics Essay Achievement Test (MEAT) adopted from 2016 West African Examination Council (WAEC) Mathematics essaytest. The items of the instrument were constructed and validated by experts in the department of examinations and quality control of the West Africa Examination Council (WAEC) and therefore requires no further validation.To determine the reliability of the instrument, the instrument was administered to 25 SS3 Mathematics students inNsukka education zone of Enugu State, Nigeria who are not part of the study population but share similar characteristics. The responses obtained from the students were duplicated and scored by three independent raters. The scores obtained from the three independent raters were subjected to Kendall coefficient of concordance (W) and the reliability coefficient of 0.91 was obtained

To collect pertinent data for the study, the instrument was administered to the SS3 students in the sampled schools by the researchers. At the end of the test, the students' responses were collected and duplicated into the number of Mathematics teachers sampled for the study as raters.The Mathematics teachers used for the study were trained before the commencement of the scoring of students' responses. Data collected were analyze usingEduG version 6.1-e to determine variance error components.

Results and Discussion of Findings

Research Question One

What are the contributions of the facets; questions, raters and interactions among the facets to the measurement errors in WAEC essay test scores in Mathematics?

Table 1: Analysis of variance component estimates for students, questions, raters and interaction on mathematics essay test

Source	Variance Component	Sum of Square	df	MeanSquare	%of Variance
S	σ_S^2	523280.0231	311	1682.5724	72.2
Q	σ_q^2	0.0000	4	0.0000	0.0
R	σ_r^2	362.5321	2	181.2660	0.1
SQ	σ_{sq}^2	85741.0667	1244	68.9237	9.3
SR	σ_{sr}^2	16293.4679	622	26.1953	0.0
QR	σ_{qr}^2	0.0000	8	0.0000	0.0
SQR	σ_{sqr}^2	68027.3333	2488	27.3422	18.4
Total		693704.4231	4679		100%

Result in Table 1 shows the seven variance components estimated for the study, which included the variance components associated with the students [σ^2(S)], question [σ^2 (Q)], rater [σ^2 (R)], student-by-question interaction [σ^2 (SQ)], student-by-rater interaction [σ^2 (SR)], question-by-rater interaction [σ^2 (QR)], and student-by-question-by-rating interaction [σ^2 (SQR)] interaction. Of the seven G-study variance components estimated, the largest variance component was the student which accounted for 72.2% of the total variance in the G-study followed by student-by-question-by-rater interaction component which recorded 18.4 % of the total variance in the G-study. However, variation due to student-by-question interaction was the third largest variance component explaining 9.3% of the total variance. Rater component recorded a very small error variance of 0.1% of the total variance. Variances due to question [σ^2 (Q)], student-by-rater and variance due to question-by-rater interaction [σ^2 (QR)] were zero (0.00) each.

The findings of the study revealed that the largest source of variation recorded is the variance associated with the student [σ^2(S)] which accounted for 72.3% of the total variance in the G-study. This shows that students' performances across task (item) are inconsistent. The finding of the study is in agreement with the findings of Lee (2005) who reported that the largest variance component was that of the examinees (students). In contrast, the findings of this study disagree with the findings of Egbulefu (2013) and Ikeh et al (2021) whose study reported that the largest contribution of error variance obtained in this study was from the residual factor, that is, student by item by rater factor followed by the students' factor. Also in disagreement was the findings of Shavelson and Webb (2005) who in their study reported that student –by-item-by -occasion interaction was the largest source of measurement error in performance assessment.

The findings of the study also reported that the second largest variance component was that of the student-by-question (item)-by-rater [σ^2 (SQR)] interaction component which recorded 18.4 % of the total variance in the G-study. The finding of the study is in agreement with the findings of Lee (2005) whose study reported that the second factor contributing to error variability in the measurement procedure is the residual. Moreover, the study disagrees with the study of Egbulefu (2013) who reported student as the second contributor to measurement error. Also, variation due to student-by-question [σ^2 (SQ)] interaction was the third largest variance component explaining 9.3% of the total variance. This showed that a significant portion of the questions were not answered by the students and that raters are consistent across the items. More so, Rater component recorded a very small error variance of

0.1% of the total variance. The findings of Kim and Wilson (2009) who reported little error variance of 0.9 of the total variance is in agreement with the findings of this study. Variances due to question [σ^2 (Q)], student-by-rater and variance due to question-by-rater interaction [σ^2 (QR)] were zero (0.00) each, suggesting that error does not emanate from any of the sources to alter rating of the students across the items.

Research Question Two

What are the differences in reliability coefficient in Mathematics essay test by increasing the number of conditions in each facet?

Table 2: Differences in reliability coefficient in Mathematics essay test by increasing the number of conditions in each facet

Facets	Initial Condition	Optimization				
Number of Raters	6	7	8	9	10	11
Reliability Estimate (Φ or Phi)	0.70	0.76	0.79	0.81	0.83	0.85
Number of Items	5	6	7	8	9	10
Reliability Estimate (Φ or Phi)	0.70	0.76	0.83	0.89	0.91	0.96

The result in Table 2 shows the differences in the reliability coefficients of Mathematics essay test that result from increasing the number of conditions in each facets of items and raters. The increase in both levels of items and raters showed a steady and gradual increase in the reliability estimate. When the level of raters was increased from 6 to 7, a reliability index of 0.76 was obtained. Moreover, a more reliability index of 0.85 was obtained from increasing further the number of raters from 6 to 11 thereby obtaining a reliability index difference of 0.14. Secondly, the same steady and gradual increase was also noticed from increasing the number of items as well. A reliability index of 0.76 was recorded from increasing the items from 5 to 6 while a more reliability index of 0.96 was obtained by further increasing the number of items from 5 to 10 thereby recording a reliability index difference of 0.25. The result shows that increasing the number of items on the score dependability was relatively large. That is, increasing the number of items produces a more generalizability coefficient than increasing the number of raters.

The finding of the study is inconsistent with the findings of Lee (2005) who reported that the impact of increasing the number of tasks on the score dependability was relatively large, but the relative impact of the number of ratings per speech sample on the score reliability was very small. The finding was in agreement with the finding of Kim and Wilson

(2009) who also reported that the effect of raters was less than that of the item in improving both generalizability (G) and reliability (Φ) coefficient. Also in agreement with the findings of the study were the findings of Guler and Gelbal (2010), Egbulefu (2013) and Ikeh (2016) who in their studies reported that increasing the number of items produces a better G and Φ coefficients.

Recommendations

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

- Test item writers should endeavor to increase the number of items of their questions to obtain a better reliability coefficient of their items.
- The researchers should embark on generalizability analysis when dealing with multiple error sources so as to estimate and disentangle error variations that contribute to measurement error and hence increases reliability and dependability of the measurement instruments.
- Examination bodies should also subject their students' scores to generalizability analysis which will help in estimating multiple sources of variations that contribute to error and determine which source affect most the performance of their students.

Conclusion

This study applies GT in estimating WAEC mathematics test scores over multiple facets. From the related literature, concepts of essay, assessment, overview of reliability, facets and sources of measurement error essay assessment were discussed. Theoretical review and related empirical studies were also discussed. The design of the study was a random effects, two-facets fully crossed s×q×r design for a G-study and D-study. The population of the study was sixty four(64) Mathematics teachers who were used as raters in scoring the instrument and eight thousand nine hundred and seventy six (8976) senior secondary three (SS 3) Mathematics students in the fifty four (54) secondary schools in Udi Education Zone. The sample size of the study was eight hundred and ninety eight (898) SS2 Mathematics students. This study involved two stage sampling, the simple random sampling technique and proportionate stratified random sampling technique respectively. Mathematics Achievement Essay Test (MAET) was used to collect data for the study. The instrument was validated by three experts, one in Mathematics and two in Measurement and evaluation unit of science education, all in University of Nigeria, Nsukka. The reliability of the instrument was established using Kendal coefficient of concordance (W) to obtain the scorer reliability of the instrument and the value of 0.719 was obtained. A computer program EduG version 6.1-e based on the Analysis of Variance (ANOVA) and Generalizability Theory (GT) was used to carry out the Generalizability analysis. Also, it was used to answer the four (2) research questions. The findings of the study are:

- The largest contribution of error variance obtained was from the students [σ^2 (S)] 72.2 % of the total variance followed by residual factor, that is, student-by-question-by-rater factor [σ^2 (SQR)] which account for 18.4% of the total variance. The third largest variance component contributing to error variance was found to be associated with the student-by-question interaction [σ^2 (SQ)] which is 9.3% followed by rater 0.1% which is variance of the rater [σ^2 (R)]. The variances due to question [σ^2 (Q)], student-by-rater and variance due to question-by-rater interaction [σ^2 (QR)] were zero (0.00) each

References

Abakpa, B. O. &Iji, C. O. (2011). Effect of mastering learning approach on senior secondary school students' achievement in Geometry. *Journal of the Science Teachers Association of Nigeria, 46* (1).

Adeyemi, B. (2010). Teacher related factors as correlates of pupils achievement in social studies in South West Nigeria. *Electronic J. Res.Edu.Psych. 8* (1): 313 – 332.

Agomuoh, P. C. &Nzewi, U. M. (2003). Effects of videotaped instruction on secondary school students' achievement in physics. *Journal of the Science Teachers Association of Nigeria 38* (1&2), 88-93

Anthony, G., &Walshaw, M. (2009). Characteristics of effective teaching of Mathematics: A view from the West. *Journal of Mathematics Education, 2* (2), 147-164.

Asikhia, O. A. (2013). Effect of cognitive restructuring on the reduction of mathematics anxiety among senior secondary school students in Ogun state, Nigeria. *International Journal of Education and Research.* 2(2) 1-20

Brennan, R. L. (2001). *Generalizability theory.* New York, NY: Springer-Verlag.

Egbulefu C.A. (2013), *Estimating measurement error and score dependability in examination using generalizability theory.* (Unpublished Ph.D thesis), University of Nigeria Nsukka.

Elaine, J. H. (2013). *What is mathematics*? Retrieved from https://www.livescience.com/38936-mathematics.html

Enu, J., Osei,K. A. &Nkum, D. (2015).Factors Influencing Students' Mathematics Performance in some selected Colleges of Education in Ghana.*International Journal of Education Learning and Development*, *3* (3) 68-74. Retrieved from https://www.eajournals.org/wp-content/uploads/Factors-Influencing-Students----Mathematics-Performance-In-Some-Selected-Colleges-Of-Education-In-Ghana.pdf

Eraikhuemen, L. (2003). The influence of gender and school location on students' academic achievement in senior secondary school Mathematics. *ife Journalof Theory and Research in Education. 7(2),* 99-112

Ikeh, F. E. (2016). Estimating multiple sources of variation and score reliability in Economics essay test using generalizability theory (Unpublished Ph.D Thesis), University of Nigeria, Nsukka

Ikeh, F. E., Ani, M. F., Kalu, I, A., Iketaku, R. I., Eze, B. A., Madu, B. C., Ene, C. U. (2021). Application of generalizability theory in estimating score dependability of economics essay test. *International Journal of Mechanical and Production Engineering Research and Development, 11* (3) 221-228

Kalijah, M.S. (2002). Education, training and careers in Physics for women in Malaysia. *IUPAP International Conference on Women in Physics UNESCO*. Paris France.

Kim, S. C & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generaalizability theory. *Journal of Applied Measurement*, 10(4) 408-423.

Kusmaryono I. (2014). The importance of mathematical power in mathematics learning. *International Conference on Mathematics, Science, and Education*, 35-40. Retrieved from https://icmseunnes.com/2015/wp-content/uploads/2015/10/7.pdf

Lee, Y. (2005). *Dependability of scores for a new ESL speaking test: Evaluating prototype tasks*. Retrieved from www.ets.org/research/policy.../imdp

Li, M., Shavelson, R. J., Yin, Y & Wiley, E. W. (2015) Generalizability theory. Retrieved from https://www.researchgate.net/publication/313966026_Generalizability_Theory

Lin, C. (2014). *Issues and challenges in current generalizability theory applications in rated measurement*. Retrieved from https://core.ac.uk/download/pdf/29153028.pdf

Mohamed, L. &Waheed, H. (2011).secondary students' attitude towards mathematics in a selected school of Maldives. *International Journal of Humanities and Social Science, 1* (15) 277-281

Myford, C. M., & Wolfe, E. W. (2002). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4), 386-422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, 5(2), 189-227.

Obioma, G. O. (2006). Emerging issues in mathematics education in Nigeria with emphasis on the strategies for effective teaching and learning of word problems and algebraic expressions. *Journal of Issues on Mathematics*, *8* (1) 1-8.

Semmelroth, C.L. (2013). Using generalizability theory to measure sources of variance on a special education teacher observation tool. Retrieved from https://core.ac.uk/download/pdf/61729504.pdf

Shavelson, R. J. & Webb, N. M. (2005). Generalizability theory: Overview. Retrieved from https://www.researchgate.net/publication/227580118_Generalizability_Theory_Overview

Soureshjani, K. H. (2011). Item sequence on test performance: Easy items first? *Language Testing in Asia*, *1* (3) 46-59. Retrieved from https://link.springer.com/content/pdf/10.1186/2229-0443-1-3-46.pdf

Uzun1 N. B., Aktaş, M., Aşiret, S &Yorulmaz, S. (2018). Using generalizability theory to assess the score reliability o communication skills of dentistry students. *Asian Journal of Education and Training,4* (2), 85-90

Valt, M.V & Maree, K. (2007). *South African Journal of Education*, 27 (2) 223-241

Yadav, D. K. (2017). Exact definition of mathematics. *International Research Journal of Mathematics, Engineering and IT, 4* (1) 34-42.Retrieved from https://1library.net/document/qo3nn7jq-exact-definition-of-mathematics.html

Yara, P.O. & Otieno, K.O. (2010). Teaching/Learning Resources and Academic Performance in Mathematics in Secondary Schools in Bondo District of Kenya, *Asian Social Science*, 6 (12) 126-132.