ESUT Journal of Education (EJE)

Vol. 6 Issue 1, May 2023

Application of Generalizability Theory in Estimating the Sources of Error in WAEC Mathematics Test Score.

[1]Ohagwu Sylvanus Chukwujekwu & [2]Ene Catherine U.

[1&2]Department of Science Education, Faculty of Education, University of Nigeria, Nsukka.

Email: [1]ohagwusylvanus@yahoo.com & [2]catherine.ene@unn.edu.ng

ABSTRACT

The objective of studying mathematics in all levels of education in Nigeria deserves harnessing the benefits and estimates sources of error in the measurement situation. This study applied generalizability theory in estimating errors in WAEC mathematics test scores. Fully crossed $s \times q \times r$ design was used. Mathematics Essay Test was used for data collection and Kendal coefficient of concordance was used in testing the reliability and 0.719 was obtained as the reliability estimate. Edu G version 6.1 was used for data analysis and the hypothesis was tested at 0.05 level of significance using standard error. The study showed that there are other sources of errors in test items. We also observed that there was an overlap of their collective variance components in both relative and absolute error variances which shows that there was no significant difference of test scores over the facets and their interactions at 0.05 level of significance. Also, the work recommended that a wider geographical scope, different D-studies, greater number of raters and items and other examination bodies should be studied.

Keywords: Mathematics, Generalizability theory, classical test theory, measurement error, decision study, reliability and facets.

INTRODUCTION

Mathematics is a subject studied in every level of education in Nigeria and made compulsory in their primary and post primary schools. Mathematics triggers the development of logical ability and sense of reasoning in an individual. Also, the knowledge of mathematics helps to figure out the foundation of ultimate development in science and industrial research. Agwagah and Gimba (2012) opines that mathematics is the foundation of all sciences, technology and even the modern developing nations depend on mathematics for survival. Moreover, most of the complicated networking and constructions done in this present world applies the knowledge of mathematics. Therefore, any source of measurement error in mathematics needs to be estimated.

Justice, Osei and Daniel (2015) opines that mathematics is seen by society as the knowledge that is necessary for the development of scientific and technological nation. Also, mathematics is one of the disciplines that does not confer to a generally acceptable definition, even among scholars. Ziegler and Loos (2017) stated that mathematics is a science that developed from the investigation of geometric figures and computations with numbers. It is important to state that a comprehensive definition of mathematics should reflect all the branches of mathematics. The inability of mathematicians to capture all branches of the discipline has resulted in different definitions given to mathematics. Furthermore, mathematics is an essential knowledge needed for the achievement of

scientific and technological nation (Anaduka& Okafor, 2013). Operationally, the researcher defines mathematics as a discipline which parades in all the activities of human endeavor and employs logic and algebra to foster creative thinking in an individual and ease some difficulties of life.

A closer examination of the objectives of studying mathematics in all levels of education in Nigeria needs to harness the benefits. These objectives include: alignment of the compulsory cross cutting subjects such as General Mathematics, English language, Trade/Entrepreneurship Studies, Civic Education; alignment of curriculum into four distinct fields of study (Nigerian educational research and development council (NERDC, 2012). Despite these importance and objectives of studying mathematics, students' scores in internal and external examinations still remain poor (Chief examiners report WAEC, 2016). No wonder while Lassa (2012) lamented that if the testers' efficiency in mathematics is not improved, the state of the development in science and technology in Nigeria will be a dream. Also, the type of education given to Nigerian children has become a source of concern due to measurement errors from sources and something vital in mathematics must be done to move the country ahead (Nalah& Daniel, 2014).Moreover, when these measurement errors are observed, generalizability theory is employed which is the statistical framework for analyzing error from various sources to produce a reliable result. Generalizability theory is more accurate when multiple error sources are involved.

In an attempt to do this, some decisions must be considered which will lead to the expected outcome of scores of students in mathematics. Also, the decision required (D-Study) is the one which uses the variance component information provided by a generalizability study (G-Study) to fashion measurement pattern that reduces error for a purpose. Decision (D) study also deals with the practical application of measurement procedure. The estimations of D-study evaluate universe scores with several reliability and dependability indices (Shavelson& Webb, 2005). In D-study, decisions rely mostly on the average over multiple observations of test items than observation of a single test item.

Also, D study are of two types: relative and absolute decision. In relative decision, grading of students are based on ranking order such as norm-referenced interpretation of test scores (Shavelson& Webb, 2005). Also the variance of errors for relative decision can be calculated mathematically using the formula below:

$\sigma^2{}_\delta = EPE\text{i}\ \ \delta^2\text{pi} = \delta^2\ \text{pie} = \frac{\sigma^2{}_{pie}}{n^i{}_i}$

In absolute decision, individuals level of performance is independent of any other performance. It is also called criterion-or domain-referenced (Shavelson& Webb, 2005). Also in absolute decisions, items main effect do not influence absolute performance because measurement errors are defined. The error variance for absolute decisions can be calculated with the formula below;

$\sigma^2\Delta = EPE\text{i}\Delta^2{}_{pi} = \sigma^2\text{i} + \sigma^2\text{pi,e} = \frac{\sigma^2{}_i}{n^i{}_i} + \frac{\sigma^2{}_{pie}}{n^i{}_i}$

The error variances are obtained from the contributions of the facets to measurement situation. More so, Facets are those factors that constantly influence every of our observed measurements. To obtain a true score variance (T) and error variance (E) in a measurement, there is a need to find as many of the facets that are needed to be observed. Hence, the universe of admissible observations (UAO) is discussed in terms of measurement facets that represent the population of an object of measurement (Brennan, 2001).

Variance of multiple facets is identified and estimate separately as well as their interaction using ANOVA. Also, facets can be fixed or random, crossed or nested.

In Random and Fixed Facets, the conditions can be exchanged with another condition from the same facet both infinite or finite number of factors in the universe of admissible observation (Ikeh, 2016). Consequently, fixed facet is synonymous to fixed factor in ANOVA. Fixed facet occurred when the decision makers used every forms of the facet in a G-study.

Crossed and Nested Facet in the other way round is when every forms of the facet are observed with all forms of other source of variation. It is denoted by $p \times i \times t$design. In a G-study design, one facet can be nested within another. This occurred when two conditions of a single facet are same to another facet (Shavelson& Webb, 2005). For example, questions in a test may be nested within the rating criteria and crossed with examinee (P). The notation of this design is P x (q: r), where r represents the facet subtests. Therefore, this study used crossed facet design of GT to estimate measurement error in WAEC mathematics test scores over multiple facets. Measurement error is the difference that exists in the measurement system when real value is subtracted from the ideal value of a measurement. Mathematically, E (measurement error) = X(ideal value) – T (real value). This shows that when the reliability of the measurement is decreased, the error increases and the real value will become high which affects mathematics test scores over the multiple facets. Linn and Burton as cited in Ikeh (2016), noted that measurement error can be used in making decisions concerning students' scores in examinations with a certain level of confidence.
Hence, the need for estimating error of students arises due to poor scores in WAEC examination and to increase its reliability. According to Nworgu (2015), reliability is the degree of consistency with which an instrument measure whatever it measures. Furthermore, every data collected for a study must contain adequate information that allows the exposure of important relationships that was hidden (Wagemaker, 2010).Traditionally, methods of reliability that are based on Classical Test Theory (CTT) have been in use which consider only one source of measurement error at a time and hence offers insufficient information about a multiple facet. For instance, test-retest reliability among others.

Some of these measurement errors that contribute to score variability include: raters, items, teachers, learners among others. These facets could influence WAEC mathematics test scores over multiple facets. Once the measurement error due to these facets are observed, the statistical frame which has a mechanism of uniting these three reliability procedures into a single estimation will be employed which calls for the use of generalizability theory (GT).
Quansah (2020) opines that generalizability theory (GT) is a statistical measurement theory which expands classical test theory to include other sources of measurement error and links the operations to the purpose of measurement.

Moreover, GT helps the researchers to separate differences in measurement and the universe the researcher draws inferences (Ene, 2015). Therefore, the researcher sees GT as statistical framework used to analyze measurement errors from various sources to produce a reliable result.

Furthermore, in an attempt to establish the WAEC mathematics test scores of secondary school students, there are facets other than students which are considered in making relative and absolute decisions concerning the scores made by the students

in examinations. These features are called sources of error such as test items; raters among others. These contribute to errors in measurement of students' test scores and affect score reliability of the measurements. Hence, it is important to identify the respective contributions of these multiple facets to measurement errors in WAEC mathematics test scores. To estimate measurement errors that influence students' test scores involves a multifaceted approach that deals with multiple sources of error which then invalidate the use of CTT that considers only one source of measurement error.

Generalizability theory is more accurate when multiple error sources are involved. The observed scores in examinations are affected by facets other than students (learners). These facets like test items, raters, among others are likely to affect reliability of WAEC mathematics test scores. The effect of these facets leads to ask about the exactness of scores obtained in schools as they were used in predicting the student's future in terms relative and absolute sense. Therefore, there is a need to find out and estimate the WAEC mathematics test scores using GT. Hence, the problem of this study is what are the relative and absolute error variances of the facets and their differences in reliability coefficient on WAEC mathematics scores?

Purpose of the Study

The general purpose of this study was to apply generalizability theory in estimating the sources of error that influence WAEC mathematics students test scores over multiple sources (facets). The study was designed specifically to obtain:

1. the differences in reliability coefficient of mathematics essay test by increasing the number of conditions in each facet.
2. the relative and absolute error variances of the facets in mathematics essay test.

Research Questions

Two research questions were formulated to guide the study.

1. What are the differences in reliability coefficient of mathematics essay test by increasing the number of conditions in each facet?
2. What are the relative and absolute error variances of the facets in mathematics essay test?

Hypotheses

A single null hypothesis was formulated to guide the study and was tested at 0.05 level of significance.

Ho1: The contributions of questions, scorers (raters) and their interactions to measurement errors in WAEC mathematics test scores are not statistically significant.

Methodology

The research design was a random effect two-facets fully crossed s×q×r design. It was used to estimate all variance components that exist in the measurement. The researcher used this design because the universe of admissible observations includes all the possible combinations of the facets. The area of the study was Udi Education Zone of Enugu State which was made up of Udi Local Government and Ezeagu Local Government areas. The population for the study was sixty-four (64) mathematics teachers as raters and eight thousand nine hundred and seventy-six (8976) mathematics students ((SS 3)) in the fifty-four (54) government owned secondary schools in the Education Zone according to its Statistic (2018/2019). The sample was eight hundred and ninety-eight (898) students and 6 mathematics teachers. This is ten percent (10%) of both students and teachers' population. Also, 27 public secondary schools which is fifty (50%) percent of the 54 secondary schools was used for the study.

Moreover, this study was sampled in two stages. In the first stage, the researcher sample the schools using simple random sampling techniques and proportionate stratified random sampling techniques for SS3 students to be included in the study as the second stage.

The instrument for the test was mathematics achievement essay test (MAET) which was adopted from WAEC questions. The items were drawn from the questions considering the five major mathematics topics that were included in the WAEC examination years (2013 – 2016) academic sessions. The data was administered and collected by the researcher with the help of the mathematics teachers. The data collected was analyzed using EduG version 6.1. Also, the hypotheses were tested at 95% confidence interval using standard error.

Result

Research Questions one

What are the differences in reliability coefficient in mathematics essay test which occur as a result of increasing the number of conditions in each facet?

Table 1: The generalizability theory analysis of the differences in reliability coefficient in mathematics essay test by increasing the number of conditions in each facet.

	Initial Condition	0ptimization				
Number of raters	5	6	7	8	9	10
Reliability Estimate (Φ or Phi)	0.76	0.76	0.79	0.81	0.83	0.85
Number of Items	5	6	7	8	9	10
Reliability Estimate (Φ or Phi)	0.76	0.76	0.89	0.89	0.91	0.96

Result in Table 1 shows the differences in the reliability coefficients of mathematics essay test that result from increasing the number of conditions in both items and raters. The increase in both levels of items and raters showed a steady and gradual increase. When the level of raters was increased from 5 to 6, a reliability index of 0.76 was obtained and 0.85 when increase from 5 to 10 thereby recording 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. Also, reliability index of 0.76 was recorded from increasing from 5 to 6 and 0.96 from 5 to 10 thereby recording a reliability index difference of 0.25. The result shows that increasing the number of items produces a better generalizability coefficient than increasing the number of raters.

Research Question two

What are the relative and absolute error variances of the facets in mathematics essay test?

Table 2: G-study of the relative and absolute error variances of the facets in mathematics essay test

Source of Variance	Differentiation Variance	Source of Variance	Relative Error Variance	% Relative	AbsoluteError Variance	% Absolute
S	107.6530		.....		.....	
	.....	Q	.....		(0.0000)	0.0
	.....	R	.....		0.0390	0.8
	.....	SQ	2.7721	60.3	2.7721	59.8
	.....	SR	(0.0000)	0.0	(0.0000)	0.0
	.....	QR	.....		(0.0000)	0.0
	.....	SQR	1.8228	39.7	1.8228	39.3
Sum of Variances	107.6530		4.5949	100%	4.6339	100%
Standard Deviation	10.3756		Relative SE: 2.1436		Absolute SE: 2.1526	

Result in Table 2 revealed that item and rater facets had an absolute variance of 0.0000 and 0.0390 respectively. Student-by-item interaction each recorded relative and absolute error variance of 2.7721. Relative and absolute error variance of 0.0000 and 0.0000 was obtained for student-by-rater interaction. Similarly, (0.0000) absolute error variance was also obtained for item-by rater interaction. The residual interaction recorded a relative and absolute error variance of 1.8228 each. Nevertheless, the overall relative and absolute error variance in making relative and absolute decisions about the students are 2.1436 and 2.1526 respectively.

Discussion

The results of this study shows that various sources of measurement errors exist in test items. The findings of this study also agrees with the findings of Ikeh (2016) and Guler and Gelbal (2010) who posits that increasing the items than raters will give a better reliability (Φ or Phi) coefficient. Also, the examination bodies, curriculum planners and researchers will know that errors exist in every measurement and to be circumspect on how to ascertain their facts during generalizations. More so, teacher factors among others were not considered which can induce score variations from different schools and may have affected the result of the study. Moreso, in testing the hypothesis at 0.05 level of significance using standard error, we observed that there was an overlap of their collective variance components in both relative and absolute error variances which shows that there was no significant difference of test scores over the facets and their interactions at 0.05 level of significance.

Conclusion

Based on the findings of the study, the following conclusions were drawn.

Increasing the number of items produces a higher reliability and generalizability coefficient than increasing the number of raters. We also observed that there was an overlap of their collective variance components in both relative and absolute error variances which shows that there was no significant difference of test scores over the facets and their interactions at 0.05 level of significance.

Recommendations

Based on the limitations and findings of the study, the researcher suggests that the work should be replicated using more geographical scope; different D-studies and greater number of raters and items. Also, the same number of raters should be used on Joint Admission and Matriculation Board (JAMB) and other examination bodies. Furthermore, Test item writers, researchers and Examination bodies should embark on generalizability analysis when dealing with multiple sources of error.

REFERENCES

Abdullahi, O. E. (2013) Inter-Relationship Between Personal Factor and Academic Achievement in Mathematics. Ebira Secondary School Students in Kogi State. *Ife PsycologicAn Inter-national Journal,* 5(1), 154-155. Retrieved June 15, 2013 fromwww.unilorin.ed.ng/publication /abudullahioe/interration between personal factor academic achievement.pdf)

Anaduka, U.S. and Okafor, C.F. (2013). *Poor performance of Nigerian students in mathematics in senior secondary examination, (SSCE);* what is not working (on line). Ret rival fromwww.agol.info/journals/iorinal.

Brennan, R. I (2001). Generalizability theory. New York Springer.

Caldwell, D. J; Sampognaro, I. & Pate, A. N. (2015). Effect of question formats on student and item performance. *American Journal of Pharmaceutical Education, 79(6); 87.*

Cronbach, I.J; Rajarantnam, N; Nanda, H & Gleser, G.C. (1972). *The Dependability of Behavioral Measurement:* Theory of Generalizability. New York, John Wiley.

Egbulefu, C.A. (2013). *Estimating Measurement Error and Score Dependability in ExaminationUsing Generalizability Theory.* An unpublished thesis. University of Nigeria Nsukka.

Elaine, J.H. (2013). *Definition of mathematics* [on line]. retrieved 15 August from http://M. linescience.com 138.

Ene, C.U. (2015). Comparison of three test theories in measurement. *Journal of general studies 2(1), 53-61.*

Gimba, R. W. &Agwagah U. N. V. (2012). Importance of Mathematics to Science and Technology. *Journal of Science, Technology, Mathematics and Education (JOSTMED) Volume 8(2).*.

Han, T. (2013). *The Impart of Rating Methods and Rater Training on the Variability and Reliability of EFL Students' Classroom Based Writing Assessment in Turkish Universities. An Investigation of Problems and Solutions.* (Unpublished doctorate dissertation) Turkey: Atatilrk University.

Ikeh, E.F. (2016). *Estimating Multiple Sources of Variation and Score Dependability in Economics Essay Test Using Generalizability Theory*. An unpublished thesis. University of Nigeria Nsukka.

Justice, E; Osei, K.A, & Daniel, N. (2015). Factors Influencing Students Mathematics

Performance in Some Selected Colleges of Education in Ghana. *An International Journal of Education Research Learning and Development, vol 3, (3)68-74*

Lassa, P.M. (2012). *The teaching of mathematics for Nigerian secondary schools,* Fab Anieh.

Lewis, W.D. and Young, T.V. (2013). The Politics of Accountability: Teacher Education Policy.Education Policy,

27(2), 190-216.doi:10.1177/0895904812472725

Lombardi, A; Seburn, M; Conley, D. & Snow, E. (2010). *A Generalizability of Cognitive Demand and Rigor Ratings of Items and Standards in an Alignment Study*. Presented at the annual conference of American educational research association. Denver, co. educational empowerment Centre. 720E.13th Ave, suite 202.

Murunga, F; Kilaha, K. & Wanyonyi, D. (2013). Emerging Issue in Secondary School Educationin Kenya. *Int. J. Adv. Res*, 1(3): 231-240.

Nalah, A.B. and Daniel, L.I. (2014). *Discourse Journal Of Education Research.* Factors Mitigating Ducational Stability of Student in central Nigerian and the way. ISSBN; 2346-7045, vol 2(1); 1-5.

NERDC (2012). *The Senior Secondary Education Curriculum Structure.* Unpublished position paper from the curriculum Centre, NERDC. Nigerian limited Jos, e-mail: fabniehpress@yahoo.com

Nworgu, B.G. (2015). *Educational Measurement & Evaluation, Theory and Practice.*Nsukka, University trust publishers.

Quansah, F. (2020).Students' Evaluation of the Quality of Teaching Using Generalizability Theory: A Case of a Selected University in Ghana *South African Journal of Higher Education. Doi: 10.20853/34-5-4212.* Retrieved from https://www.semanticscholar.org.s...

Shavelson, R.J. & Webb, N.M. (2005). Generalizability theory. 1973-1980. *British journal of mathematics and statistical psychology, 34, 133-166.* http//dx. Doi.org/10,1111/j,2044-8317. Tb00625.x. *Journal of Psychology, 15, 72-101.*

Wagemaker, H. (2010). IEA; *International Studies, Impact and Transition*. Paper presented in the 4th IEA international research conference. July1-3. Gothenburg, Sweden. Retrieved from http;//www.iea.u/irc-2010.htmL.

West African Examination Council (2016). Chief Examiner's Report. Lagos: WAEC.

Ziegler, G.M., Loos, A. (2017). "What is Mathematics?" and why we should ask, where one should experience and learn that, and how to teach it.13th International Congress on Mathematical Education. ICME-13 Monographs. Springer, Cham. https://doi.org/10.1007/978-3-319-62597-3_5