British Journal of Contemporary Education
Volume 2, Issue 1, 2022 (pp. 64-73)
www.abjournals.org

ASSESSING SCORE DEPENDABILITY OF WEST AFRICA EXAMINATION COUNCIL (WAEC) 2019 MATHEMATICS OBJECTIVE TEST USING GENERALISABILITY THEORY

Imasuen Kennedy[1] and Stanley O. Ebuwa[2]

[1]Institute of Education, University of Benin, Nigeria

[2]ICTU Department, University of Benin, Nigeria.

Email: kennedy.Imasuen@uniben.edu[1]; zustan@yahoo.com[2]

Tel: +2348109670163

Cite this article:

Imasuen K., Stanley O.E. (2022), Assessing Score Dependability of West Africa Examination Council (WAEC) 2019 Mathematics Objective Test Using Generalisability Theory. British Journal of Contemporary Education 2(1), 64-73. DOI: 10.52589/BJCE-OCA9OZJT

Manuscript History

Received: 14 July 2022

Accepted: 12 Aug 2022

Published: 11 Sept 2022

ABSTRACT: *This study investigated score dependability in the 2019 West Africa Examination Council (WAEC) Senior Secondary School examination using the generalisability theory. The study was specifically concerned with identifying and analysing the score dependability of the Senior Secondary School 2019 WAEC mathematics objective examination using generalisability theory, and determining the highest contribution of facets: students, items and teachers to score dependability. Two research questions were raised to guide the study. The study was a survey which adopted a random effect two-facet fully crossed* $s \times r \times i$ *design for generalisability (G) and decision (D) studies. The population consisted of fifty-six thousand, seven hundred and ninety-seven (5697) Senior Secondary three (SS3) students in the seventy-five (75) public secondary schools in Benin Metropolis for the 2019/2020 academic session. The instrument for data collection was a fifty (50) multiple choice WAEC, Mathematics 2019 examination. The instrument had been validated by the West African Examination Council (WAEC). The reliability of the items was ascertained using the Kuder – Richardson 20 (KR 20) to obtain internal consistency. It gave a value of 0.92. Data collected were analysed using the software EduG version 6.0-e based on analysis of variance (ANOVA) and generalisability. The findings which emerged from the study were the highest effects to score dependability in examination came from the interaction of students and teachers, an index of dependability (∅)of 0.92 high enough to maximise reliability was observed only when the teachers were increased to 78. Based on the findings, it was recommended that generalisability analysis should be carried out by researchers, test developers and examination bodies so as to reduce or eliminate measurement error and hence maximise reliability, and there should be enough invigilators when conducting examinations, thereby minimising error and maximising reliability (dependability) of examination scores.*

KEYWORDS: Reliability, Mathematics, Generalisability, Dependability, Facets.

Article DOI: 10.52589/BJCE-OCA9OZJT
DOI URL: https://doi.org/10.52589/BJCE-OCA9OZJT

INTRODUCTION

Reliability as a psychometric property of any measuring instrument deals with stability and constituency of scores when the instrument is used over time. Several authors defined reliability in various ways. For example, Kline (2000) opined that reliability with respect to tests has two distinct meanings. One refers to stability over time and the second is an internal constituency. Mcleod (2007), stated that reliability in psychological research refers to the constituency of a study or measuring test. On their part, Wilkinson and Robertson (2006), posited that reliability with respect to research means repeatability or constituency. Meyer (2010) opined that reliability is the degree to which an assessment tool produces stable and consistent results. According to the National Council in Measurement in Education (1999), reliability in statistics and psychometrics is the overall consistency of a measure. It further stated that a measure is said to have high reliability if it produces similar results under consistent conditions. On his part, Bolarinwa (2015) averred that reliability is the extent to which a questionnaire, test, observation or any measuring procedure produces the same result on repeated trials. It is also seen as the stability or consistency of scores over time (Miller, 2015). From the foregoing, it is clear that reliability means stability and consistency of scores obtained from measuring instruments over a period of time.

Kaplan and Saccuzzo (2005) stated that there are four broad types of reliability: test-retest, alternate form, internal consistency and interrater. Test-retest reliability is also known as 'test me, come again to test me'. It involves two separate administrations, usually within a space of two weeks and the two scores from the two administration is correlated using the Pearson Product Moment Correlation Coefficient. The alternate or parallel reliability is a measure of the similarity of two forms of a test. For forms to be considered parallel, they must have exactly the same difficulty level. One drawback of this type of reliability is that it is usually not achievable. Internal consistency deals with the relationship between the items, that is, if the items are related. This involves a single administration of the instrument. For interrater reliability, Sattler in Sandilos and DiParna (2011) stated that it is concerned with the constituency across different raters when assessing a behaviour, trait or construct.Reliability is a term that is frequently used in psychology, but one that differs slightly depending on the definition. Two reliability models in the literature on psychometrics are the true score model of the classical test theory (CTT) developed by Spearman in the early 1900s and the generalisability theory (GT) by Cronbach and associates in 1972. Both emphasise stability; while CTT assesses the repeatability and constituency of measures, GT focuses on the dependability or accuracy of the generalisation of the test score based on the purpose and components of the testing situation. Generalisability analysis estimates the dependability (reliability) of measures. Classical test theory is the foundation of reliability theory and stated that an individual's observed score is equal to his/her true score plus random or unsystematic error (Sattler, 2001). According to Shavelson and Webb (1991), CTT is mainly concerned with the relative standing of individuals; it assumes that a hypothetical true score exists and that the forms of an assessment were parallel.

Generalisability theory, on the other hand, is a statistical theory about the dependability of behavioural measurement (Cronbach et al in Ogunka & Orluwene (2020). It liberalises classical test theory by using analysis of variance (ANOVA) methods to untangle multiple sources of error, by the researcher that contributed to the undifferentiated error (E) in CTT. It is also a statistical theory for estimating the reliability of behavioural measurement which gives researchers ample opportunity to comprehensively assess numerous sources of measurement

Article DOI: 10.52589/BJCE-OCA9OZJT
DOI URL: https://doi.org/10.52589/BJCE-OCA9OZJT

error (variance components). GT concern itself with the relative and absolute dependability of behavioural measures. GT is a framework for analysing how well-observed scores allow users to make generalisations about a person's behaviour (Shavelson & Webb, 1999). Instead of partitioning and observing scores into two as in the case of CTT, a true score and error score without differentiating the various sources that contributed to the error is seen as a major limitation of CTT (Baykul, 2000; Guller, 2009). However, generalisability theory on its parts partitions the error variance into multiple components representing several different sources of error simultaneously and shows the contribution and influence of each. Hence, several authors such as Brennan (2001), Shavelson and Webb (1999) see generalisability theory as an extension of CTT with the addition of separating the various sources of error and estimating the contribution of each to measurement error and score dependability. Another advantage of generalisability theory as stated by Brennan (2001) is that it can estimate the reliability of mean ratings for each examinee, while simultaneously accounting for both interrater and intra-rater in consequence as well as discrepancies due to various possible interactions which are impossible in CTT.

In a generalisability theory, each source of variation, such as the items, raters, or different measurement situations available in the measurement process is called a facet. Brennan (2001) opined that facets can be interpreted as the measurement situations having similarities. Each level on the facet is called a condition, while the source revealing the variability of concern (student, items etc.) is called the object of measurement. In this study, the object of measurement is students (s), and the two facets are items (i) and teachers/raters (r). Two studies are usually conducted in a generalisability theory. They are the generalisability study (G-study) and the decision study (D-study). A G-study is carried out to ascertain how well the scores can be used for multiple situations. It involves estimating variance components that might in turn be used in a D-study for computing the generalisability coefficient. On the other hand, D-study is conducted for the purpose of optimisation. There are also two types of decisions to be made in generalisability theory; relative and absolute decisions. The relative error is analogue to the error variance in CTT (Lee & Frisbie, 1999). There are also, two reliability coefficients, the generalisability coefficient (G coefficient) and dependability index (Phi).

Generalisability theory is not based on the traditional assumption that reliability and validity are separated but assumes that reliability and validity both fall on the same continuum of dependability (Silva, in Poncy, 2006). What teachers are interested in when they administer a test is to see if that score is dependable. Inherent in this view, is that scores will differ from one administration to another due to a lot of factors which include test administration, occasion, test forms, rates and so on. It is only generalisability theory that can pinpoint and estimate these sources of errors that causes inconsistency in the generalisation of test scores.

Kin and Wilson (2009) defined dependability of behavioural measures as the accuracy of generalising from a person's observed score on a measure or a test to the score that the person who has received averaged over all possible conditions. This type of variation that is mainly due to the measuring instrument rather than factors which are directly controlled by the examinee denotes uncertainty in the quantitative description of the individual on the basis of the test.

According to Shavelson and Webb in Ogunka and Orluwene (2020), dependability refers to the accuracy of generalising from a person's observed score on a test or rather other measures (behaviour observation, opinion survey) to the average score that person would have received

Article DOI: 10.52589/BJCE-OCA9OZJT
DOI URL: https://doi.org/10.52589/BJCE-OCA9OZJT

under all the possible conditions that the test user would equally willing to accept. This notion of dependability is the assumption that the person's knowledge, attitude, skills, or other measured attribute are in a steady state; it is assumed that any differences among scores earned by the individual on different occasions of measurement are due to one or more sources of error, and not to systematic changes in the individual due to maturation or learning.

Orluwene (2020) indicated that in the measurement of complex traits imperfect instruments are used so that the score observed for each person almost always differs from the person's true ability or characteristics; she further affirmed that the discrepancies between the true ability and the observed ability results from measurement error, which implies some inaccuracy in the measurement exist because measurement error may inflate or depress any subject's score in an unpredictable or predictable manner.

The comparison of dependability of reliability in generalisability theory and classical test theory to determine standard error measurement varies. Atilla (2012) asserted that the use of classical test theory approaches to determining score reliability, however, is not capable of identifying and untangling this profusion of error which classical reliability was not conceptualized to do since it accounts for only one source of error at a time. Similarly, Ikeh and Madu cited by Tavakol and Brennan (2013) state that Classical Test Theory (CTT), assume that the student's true score is the sum of the student's observed score and a single undifferentiated error term. Kpolovie (2010) asserted that classical test theory has reliability embedded in the true score and the error score model defines reliability as the coefficient of the predictable proportion of variance in observed scores from the true scores.

Esomonu and Okeaba (2021) estimated measurement error and score dependability of the inventory for students' integration into the University Academic Culture using generalisability Theory. The results show that the highest contribution to measurement error in ISIUAC scores was the residual which accounted for 85.6% of the total variance. The analysis produced a relative standard error variance of 0.22189 which resulted in a generalisability coefficient of 0.55 and an absolute error variance of 0.23510 which resulted in a dependability coefficient (ϕ) of 0.52. The result of the D-study revealed that a minimum of 100 question items were needed to produce generalisability and dependability indices of 0.82 and 0.80 which both attained the benchmark. The variance components of the facets: students, questions, and their interactions overlapped, indicating that the variance components were not significantly different in their contributions to measurement error in ISIUAC scores.

McLaughlin, et al (2017) examined the dependability of the Learning Target Rating Scale (LTRS) using generalisability theory. The result of the study showed that the percentage of the variance of total LTRS scores accounted for by the different sources of variance in the model was similar across the three occasions, with learning targets and teachers accounting for the largest percentage of variance while raters and children accounted for a small percentage of variance. Ogidi (2021) utilised generalisability theory in the estimation of variance components in National Examination Council Essay Questions in Christian Religious Studies. Results of the study showed the index of dependability of 0.938 was obtained which indicated that the instrument was adequate for the certification examination.

Bamidele, et al (2021) carried out a study in estimating generalisability and dependability indices of students' scores in teaching practice assessment in a Nigerian College of Education. It was observed from the result obtained that the dependability coefficient/index of the

Article DOI: 10.52589/BJCE-OCA9OZJT
DOI URL: https://doi.org/10.52589/BJCE-OCA9OZJT

2016/2017 teaching practice scores; the obtained D study or dependability index was high (0.74) considering the 0.70 level of acceptability value, therefore, the dependability index of the 2016/2017 teaching practice was high. The high dependability index level of the 2016/2017 teaching practice scores may be due to the contribution of four sources of measurement errors and to the difference in the persons' performance and high level of commitment of students during the 2016/2017 teaching practice programme.

Statement of the Problem

A student's performance in a given examination is usually gauged by several characteristics other than the student's factor. These characteristics are also known as sources of error and they include test questions, invigilators, and so on and affect the score dependability of these measurements. The impact of these factors leads to questions about the accuracy, precision, and ultimately, the fairness of the scores obtained by students in any given examination. More so, scores obtained by the objects of measurement, (students) in the examination are affected by multiple sources of error and scores from the examinations are used in making relative and absolute decisions concerning students, there is the need to estimate score dependability of examinations using generalisability theory, so as to determine the contributions of each of these facets in measurement situations in examinations with a view to minimising and maximising the reliability of their scores. Estimating the score dependability of any given task involves a multifaceted approach which the classical test theory cannot address as it addresses only one source of measurement error. In the light of this, the present paper seeks to assess the score dependability of the WAEC 2019 Mathematics objective test using generalisability theory.

Research Questions

The following questions were raised to guide the study

1. What is the contribution of the facets: students(s), items (i), and raters (t) to score dependability in the WAEC 2019 Mathematics objective test?
2. To what extent do the dependability coefficients show the degree to which students maintain their rank order across facets: item (i), and raters (t) in WAEC 2019 Mathematics objective test scores?

METHODS

The study was a survey which adopted a random effect two-facet fully crossed $s \times r \times i$ design for generalisability (G) and decision (D) studies. The fully crossed design in the G – study was used to estimate all the possible variance components in the measurement situation. The D – study used the information provided by the G – study to design the best measurement procedures minimising undesirable sources of measurement error and maximising reliability. The population of the study was all the senior secondary three (SS3) students of public secondary schools in the Benin metropolis for the 2019/2020 academic session. They were considered appropriate for the study because they should have almost covered the syllabus for mathematics in any of the external examinations, have stayed six years in school and are fully prepared for any form of examination. There are four local government areas in Benin metropolis and they are Egor, Oredo, Ikpoba-Okha and Ovia North–East. There are seventy-

Article DOI: 10.52589/BJCE-OCA9OZJT
DOI URL: https://doi.org/10.52589/BJCE-OCA9OZJT

five (75) public senior secondary schools in these four local government areas with a student population of 5697 students. 570 students which represent 10% of the total population of SS3 students in the four local government areas constituted the sample. They were selected from thirty-eight (38) schools in the locality. The multi-stage sampling technique was adopted for the study.

The instrument used for data collection was a fifty (50) multiple choice of the 2019 WAEC mathematics objective questions for the 2019 examination year. The objective items were constructed by WAEC and are assumed to have been validated and standardised before it was administered to the students. The items covered a range of topics in Mathematics showing that it is also content valid and considered appropriate for utilization in the study. The reliability of the instrument was established using a sample of 50 students and five teachers from public senior secondary (SS 3) who were not used in the main study. The reliability of the instrument was determined using the Kuder – Richardson 20 (KR 20) to obtain internal consistency. It gave a value of 0.92.

Data collected were analysed using computer software, EduG version 6.0-e based on analysis of variance (ANOVA) and generalisability theory.

RESULTS

Table 1: A generalisability study showing the effects of students, teachers, items and their interactions to score dependability in 2019 WAEC examination

Sources	Variance component estimates	Relative error variance	% Relative variance	Absolute error variance	% absolute error variance
Students (s)	22.70349 ….				
Teachers (t)	0.000	.....		.....	
Items (i)	0.000	.....		(0.00000)	0.0
s × t	0.000	.....		.....	
s × i	0.000	(0.00000)	0.0	(0.00000)	0.0
t ×i	0.000	0.00166	100.0	0.00166	100.0
s × t × i	0.000	.....		(0.00000)	0.0
Total		0.00166	100%	0.00166	100%

Error Variances

$\sigma^2 \delta = 0.00166$

$\sigma^2 \Delta = 0.00166$

Coefficients

$E\rho^2 = 0.91$

$\emptyset = 0.82$

Article DOI: 10.52589/BJCE-OCA9OZJT
DOI URL: https://doi.org/10.52589/BJCE-OCA9OZJT

Table 1 showed that the absolute error variance for items, teachers, the interaction of items and students, teachers and items were set to zero. Conversely, the absolute error variance estimate for the interaction of teachers and students was 0.00166 accounting for 100% of the absolute percentage. The dependability index (∅) of 0.82, showed that 38 teachers supervising 570 students yielded a high dependability index.

Table 2: Estimated dependability coefficient (∅) for a fully crossed s × t × i D-study Design with a different number of teachers

Number of teachers	∅
38	0.74
48	0.83
58	0.86
68	0.91
78	0.92

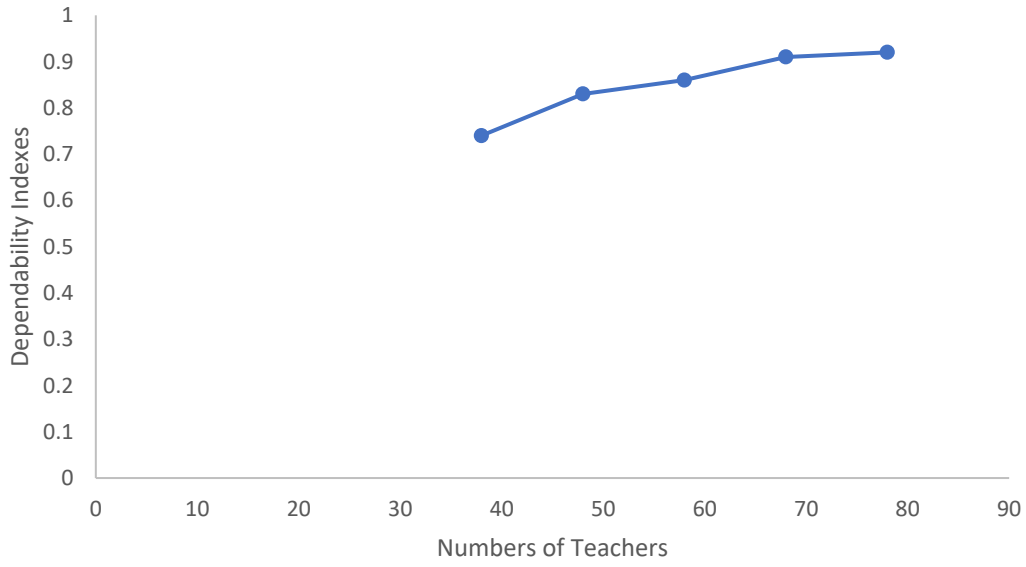

Figure 1: Dependability indices resulting from relative decisions for different teachers.

Table 2 and figure 1 showed that with 38 teachers the dependability index (∅) was 0.74. When the number of teachers was increased to 58, the dependability index (∅) was 0.86, an increase of 0.10. An increase in the number of teachers to 78 produced an increase of 0.16 in the dependability index (∅). This showed that the performance of an individual student does not affect the performance of another student.

Article DOI: 10.52589/BJCE-OCA9OZJT
DOI URL: https://doi.org/10.52589/BJCE-OCA9OZJT

DISCUSSION OF FINDINGS

The findings from the study revealed that the highest effects to score dependability in examination came from the interaction of students and teachers. Items and the interaction of students and items did not have any effect on score dependability in examination scores. This implied that the strictness of the teachers in terms of invigilation on the students maximized their observed scores in the examination. Also, it can be observed that more of the absolute error variability in the examination came from teachers (invigilators), changing the level (numbers) of teachers will have a large effect on the score dependability than changing the number of items. Therefore, there will be the need to bring in more teachers to bring about dependable scores in examinations. These findings in the study were consistent with the earlier findings of Lee et al (2001), Fulcher (2003), Ogidi (2010) and Bamidele et al (2021).

Another revelation from the study was that with a dependability index of 0.92, students that passed were comfortably separated from those that failed. Students who had attained the predefined score and above were separated from those students who did not perform well. The level of invigilators at 38 was not quite satisfactory to produce an absolute scale of measurement. There should be at least 78 teachers so as to attain a dependability index (Ø) of 0.92 that will help to successfully separate students in terms of their performance irrespective of the performance of other students. The result was consistent with the study of Brennan (2001) who found that more raters were needed for a high dependability index. The findings of the study were also supported by Lee (2006) who opined that an increase in the number of raters yielded a higher dependability index than when the raters were small in a study on the dependability of scores for a New ESL Speaking Test. It was also corroborated by Esomonu and Okeaba(2020) who revealed that a minimum of 100 question items were needed to produce dependability indices of 0.80 to attain the benchmark.

CONCLUSIONS

Generalisability theory provides an integrated framework for evaluating multiple sources of variability in examination scores and for deriving implications for test development and test scores interpretation. Apart from the student factor, other sources (facets) affect the scores students obtain in examinations. In this study, the interaction of students and teachers contributed had a large effect on score dependability in the examination. Above all, an increase in the number of the facet -teachers (invigilators) showed that a high index of dependability (Ø), was high enough to rank order student relatively.

Recommendations

Based on the findings of this study, the following recommendations were made.

- Generalisability analysis should be carried out by test developers and examination bodies in the estimation of reliability so as to estimate multiple sources of error and reduce or eliminate measurement error and hence maximise reliability.
- In generating items, item writers should endeavour to develop items that will discriminate among students of different achievement levels. This will in no small way reduce error in measurement and ensure score dependability.

Article DOI: 10.52589/BJCE-OCA9OZJT
DOI URL: https://doi.org/10.52589/BJCE-OCA9OZJT

- There should be enough invigilators when conducting examinations. This will help in reducing cheating among the object of measurement (students), thereby minimising error and maximising the reliability of examination scores.

REFERENCES

Atilla, Y. (2012). Dependability of job performance ratings according to generalisability theory. *Journal of Education and Science*, 163(37), 157 – 348.

Bamidele, S.T., Gana, A.Y., Kehinde, A., & Adekunle, A.R. (2021). Estimating generalisability and dependability indices of students' scores in teaching practice assessment in a Nigerian College of Education. *Sapientia Foundation Journal of Education, Sciences and Gender Studies (SFJESGS)*, 3 (2); 7 – 15 ISSN: 2734-2522 (Print); ISSN: 2734-2514 (Online)

Baykul, Y. (2000) *Classical Test Theory and Practice of Measurement in Education and Psychology*. OSYM Publications, Ankara, Turkey.

Bolarinwa, O.A (2015). Principles and methods of validity and reliability testing of questionnaires used in social and health science research. *Niger Postgraduate Medical Journal* 22:195-201

Brennan, R.L (2001) *generalisability Theory: Statistics for Social Science and Public Policy.* Springer-Verlag Berlin Heidlberg. New York.

Esomonu, N.P., & Okeaba, J. U. (2021) Estimating Measurement Error and Score Dependability of the Inventory for Students' Integration into the University Academic Culture (ISIUAC) Using generalisability Theory. *Rivers State University Journal of Education (RSUJOE)*, 24 (1):35-46

Fulcher, G. (2003): *Testing the second language speaking*. Harlow: Pearson Professional Education.

Güler, N. (2009). generalisability Theory and Comparison of the Results of G and D Studies Computed by SPSS and Genova Packet Programs. *Education and Science*, 34, 154.

Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). (1999). Standards for educational and psychological testing. Washington, DC: AERA.

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

Kline, R. B. (2000). Beyond significance testing: Reforming data analysis methods in behavioural research. Washington, DC: *American Psychological Association*

Kaplan, R. M. & Saccuzzo, D. P. (2005). *Psychological testing: Principles, applications, and issues (6th Ed.).* Belmont, CA: Thomson Wadsworth

Kpolovie, P. J (2010). *Advanced Research Methods*, Owerri Springfield Publisher ltd.McGrew, K. S., Johnson, D. R., Cosio, A., & Evans, J. (2003). Increasing the chance of no child being left behind: Beyond cognitive and achievement abilities. Unpublished manuscript, University of Minnesota.

Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, 23(2), 131-166.

Article DOI: 10.52589/BJCE-OCA9OZJT
DOI URL: https://doi.org/10.52589/BJCE-OCA9OZJT

Lee, G., & Frisbie, D. A. (1999). Estimating Reliability Under a generalisability Theory Model for Test Scores Composed of Testlets. *Applied Measurement in Education*. 12(3), 237-255.

Marty, M. C., Henning, M. J., & Willse, J. T. (2010). Accuracy and Reliability of Peer Assessment of Athletic Training Psychomotor Laboratory Skills. *Journal of Athletic Training*, 45(6):609–614.

McLaughlin, T. W., Snyder, P. A. & Algina, J. (2017) Using generalisability Theory to Examine the Dependability of Scores from the Learning Target Rating Scale. *Topics in Early Childhood Special Education*, 37(3) 164–175.

McLeod, S. A. (2007). *What is Reliability?* Retrieved on 27the June 2017 from www.simplypsychology.org/reliability.html

Meyer, P. (2010). Reliability: Understanding statistics measurement. New York, NY: Oxford University Press.

National Council on Measurement and Evaluation in Education (1999)

Miller, M.J (2015). *Graduate Research Methods*. Available from: <="" a="">. [Last accessed on 2015 Oct 10].

Ogidi, R.C (2021) Application of generalisability theory in the estimation of variance components in national examination council essay questions in Christian religious studies in Ogba/Egbema/Ndoni local government area of Rivers State, Nigeria. *European Journal of Research and Reflection in Educational Sciences*, 9 (2): 1-8

Ogunka, R.I & Orluwene, G. (2020). Application of generalisability theory in estimating variance components in National Examinations Council problem-solving questions in mathematics. *European International Journal of Science and Technology*, 9(4). 61-69

Orluwene, G.W (2012). *Fundamentals of testing and non-testing tools in educational psychology*. Harey publishers.

Poncy, B. C (2006). An investigation of the dependability and standard error of measurement of words read correctly per minute using curriculum-based measurement. PhD diss. The University of Tennessee. http://trace.tennessee.edu/utk_gradiss/1993.

Sandilos, L.E., & DiPerna, J.C (2011). Interrater reliability of the classroom assessment scoring system-pre-k (CLASS Pre-k). *Journal of Early Childhood and Infant Psychology*. 7: 65 -85

Shavelson, R. & Webb, N. (1991). "generalisability theory; 1973-1980; *British Journal of Mathematical and Statistical Psychology*, 34, 133-166.

Tavakol, M. & Brennan, R.L (2013). Medical education assessment: A brief overview of conce[ts in generalisability theory. **International Journal of Medical Education**.4: 221- 222

Wilkinson, G. S., & Robertson, G. J. (2006). *Manual for the Wide-Range Achievement Tes*t (WRAT–4). Los Angeles, CA: Western Psychological Services

Article DOI: 10.52589/BJCE-OCA9OZJT
DOI URL: https://doi.org/10.52589/BJCE-OCA9OZJT