SJIF Impact Factor: 6.260| ISI I.F.Value:1.241| Journal DOI: 10.36713/epra2016 ISSN: 2455-7838(Online)

EPRA International Journal of Research and Development (IJRD)

Volume: 5 | Issue: 4 | April 2020 - Peer Reviewed Journal

APPLICATION OF GENERALIZABILITY THEORY IN RELIABILITY ASSESSMENT OF CHEMISTRY ESSAY ACHIEVEMENT TEST

Orluwene, Goodness Wobiele

Department of Educational Psychology, Guidance and Counseling, Faculty of Education
University Of Port Harcourt

Memory, Queensoap

Department of Arts Education, Faculty of Education, Federal University, Otueke, Bayelsa State

ABSTRACT

The study used generalizability theory (GT) to estimate the reliability of the West African Senior School Certificate Examination in Chemistry essay questions conducted in May/June 2018. It was guided by two research questions. The study adopted a two –facet fully crossed random design (S x I x R) using a sample of 74 senior secondary three chemistry students in Obio/Akpor Local Government Area of Rivers State, Nigeria. A two-stage sampling method via simple random technique by balloting at stage one and then non-proportionate stratified random sampling and accidental/convenience sampling at stage two, were used to obtain the required sample. Data were collected using an adopted instrument tagged WASSCE Chemistry paper 2 question conducted by West African Examination Council in May/June, 2018. It is made up of two sections, section A has only one item for all candidates while section B has four items, for Nigerians, so in all it has five questions/ items. Data obtained were analysed using factorial analysis of variance by VARCOMP procedure, via univariate model. The results obtained indicated that multiple sources of error include students, items, raters, interaction effects between students and items (S1), students and raters (SR), items and raters IR and students, items and raters SIR. However,the largest component of error resulted from SI followed by SIR, item, raters, SR and then IR. Based on the findings, it was recommended among all that information obtained from WASSCE chemistry paper 2 conducted in may/June 2018 should be depend on. Again that generalizability theory should be used in determining the reliability of any given measure due to its ability to disentangle multiple sources of error.

INTRODUCTION

Chemistry is a central scientific subject that deals with the study of matter, its composition and interaction with the world. It is widely used in almost every facets of life such as food, clothing, health, shelter, career development, industries and technological aspect of the societal development. For instance chemistry is a subject that is related to many scientific disciplines such as engineering, pharmacy, medicine, agriculture laboratory scientist, biochemistry and even, teaching, technologists, nutritionists and so on. Based on all these, the importance of chemistry in life and societal developments cannot be overemphasized. Since it is glaring that everything on earth depends so much on chemistry thus, chemistry helps in our day to day decision that affects our lives.

To actually gain all the benefits in chemistry there must be higher level of chemistry achievement among students in the post primary schools. Achievements in chemistry among students in the post primary schools determines the quantity and quality of those who will further their education in studying chemistry and its related disciplines as earlier mentioned. The achievement of students in chemistry is determined by their raw scores in the subject after assessment.

In assessing students' achievement in chemistry different tools are employed. These tools could be oral test, work-sample test and sometimes paper and pencil tests which may take different item formats such as objective and essay item formats. It is observed that some examining bodies such as West African Examination Council (WAEC) used work sample test (practical in chemistry, and a combination of objective and essay questions, to achieve adequate assessment of the students/candidates skills. This is because the limitation of one item format say objective may be covered by the inclusion of essay item format. However in this study only the essay questions conducted by WAEC to assessed students' chemistry achievement in 2018 was considered.

Essay questions help to develop critical and logical thinking among students (Orluwene, 2012). They are questions used to assess the students' ability to organize and present their ideas in a logical and coherent manner. Through essay questions students can demonstrate their initiative and the originality of their thoughts and so on.

On the other hand, as it is said every coin has two sides, essay test despite its merits as earlier stated have the problem on unreliable scoring whether intra, inter-rater or on repeated measurement. This is because it is polytomously scored and students' scores are affected by the attitude of the scorer (Orluwene, 2012). On this basis different raters can come up with different scores if given opportunity to rate the students' responses independently. Sometimes the same rater can also grade the same question differently at two or more different occasions. As Anatol and Hariharan (2009) rightly stated that, the grading of essay question is challenged by subjectivity, Halo-effect and uneven variability. Coffman in Gugiu, Gugiu and Baldus (2012) stated that the variations in the scores of students may depend on the employment of different rating standards by different raters, different criteria employed for rating responses by the same rater at different occasions or by different raters independently.

In another dimension Brown (2010) reported that factors like gender, students name, how responses are organized and presented and the language used in the write-up may also influenced the variation in students' scores. In addition, difficulty levels of the essay questions may also contribute to the variation in students scores rated by different raters independently or the same rater at different occasions.

Anatol and Hariharan (2009) reported that several studies that investigated the reliability of the students' scores in essay questions rated by multiple raters indicated that reliability coefficient obtained ranged from very low to fairly high. To support the above assertion, Gugiu et al (2012) stated that the problem of low reliability of essay questions dates back to 1930 when the reliability coefficients that ranged from 0.25 to 0.51 were obtained from an instrument that 61 teachers graded within 11 weeks independently. The one graded by five instructors yielded the coefficient range of -0.41 to 0.85; for 16 instructors the coefficient range of 0.42 to 0.91 was obtained.

In all, it is presumed that, it is difficult to obtain reliable scores from the ratings on essay questions mostly when CTT is applied mean while the scores of the students in the test whether objective or essay tests represents the results obtained from assessment, which is used for prediction selection, certification, classification, placement and evaluation decisions (Orluwene, 2012).In other words, test results are useful for decision making, so for authentic decision to be made, the tests used to obtain the data that guided the decision making must be of good quality or possess good psychometric properties. Psychometric properties of a test provide good and sound information about the meaningfulness and usefulness of the test and its result. The psychometric properties of a test include validities, reliabilities, item's difficulty, discrimination and distracter levels. However, among all the psychometric properties mentioned, the present study focused on the reliability of the West African Senior School Certificate Examination in chemistry essay questions conducted in the year 2018.

Reliability of a test also known as the dependability of a test is the authentic means of generalising the correctness of students observed score on a given test to the universe score that those students would have received under different forms of a test, (equivalent forms) different testing conditions (test-retest) different raters (inter-rater) and internal consistency etc. Dependability is the accuracy of generalizing from a person's observed score on a test or other measures to the average score that students would have received over all possible testing conditions.

It entails the quantification of how a given test is consistent or inconsistent in reproducing students' observed scores in a repeated measurement (Brennan, 2011). Reliability is the extent to which a measure yields consistent results (Ary, Jacobs & Razavieh, 2002). To Anastasi and Urbina (2006:98) reliability is the consistency of scores obtained by the same person when re-examined with the same test on different occasions or with different sets of equivalent items or under other variable examining conditions. To this end, the researchers viewed reliability as the extent to which students scores are nearly the same or unchanged on repeated measurement or equivalent measures.

Reliability is one of the principal qualities of a test. It determines the level to which the decision made based on the data collected from a measure is authentic. In other words it determines the level of confidence the test users will have on the test and its results. Orluwene, (2012) asserted that an unreliable measure of a variable will not provide an accurate indication of the individual's level in that variable. This concur with Elliot, Kratochwill, Cook and Travers (2000:432) assertion that unless a test is reasonably consistent on different occasions or with different samples of the same behaviour, one can have very little or no confidence in its results. On the other hand an assessment that provides inconsistent results cannot be depended upon to provide information useful for authentic decision making. In all, if confidence is to be placed on the data obtained from any test, the results obtained from it must be highly consistent regardless of the method for quantifying the reliability of that test.

In recognition to the role of reliable instruments to decision making, the desire to design assessments, examination and tests that are free from measurement error became a big concern to most test users such as classroom teachers, examining bodies and recruitment personnel's (Rust, 2007). In justifying the credibility of the decision to be made or made, obtaining high level of assurance on the data or results from the test, the test developer employed different approaches of assessing the reliability of the test. These approaches are Classical Test Theory (CTT), Generalizability Theory (GT) and Item Response Theory (IRT) (Schuwirth & Vleuten, 2011). Classical test theory is the oldest and most used theories in establishing reliability and other qualities of a test. It is centred on the assumption that an individual's observed score is a component of true score and error score (i.e. $O_s = T_s + E_s$).

The true score represents the score a student obtained as a result of his/her ability while the error score is the score obtained as s result of any condition that is irrelevant to the purpose of the test. In classical test theory, the multiple sources of errors are not distinguished so measurement error is regarded as the undifferentiated random variation. Based on this, an individual's true indication of his/her ability in a given test may not be accurately known. The error component for items may reflect differential item difficulties or easiness while component for occasion reflects the different periods in which the test was administered to the students. The component for test forms reflects the difference in the composition of two or more different forms of test administered to students. Sometimes an observed score may have higher or more components of the error score than the true score, as such, decision that will be made from such data may be undependable. This is because measurement error is accidental deviation that is different in each individual case and occurs in parts in every direction according to the laws of probability.

It is also a randomly entangled error that may lead to increase or decrease in the observed score (Onunkwo, 2002).

Furthermore, in CTT, the reliability coefficient is expressed mathematically as reliability coefficient =

True score variance
… equation 1
True score variant + total error variance

From equation 1, it is clear that error variance in CTT is a single entangled (undifferentiated) entity. So in CTT, the partitioning of the observed score into true score and error score is likened unto using one-way analysis of variance to partition systematic and random error effects. Shavelison and webb (2005) asserted that with the undifferentiated measurement error results obtained cannot be generalised.

Indeed, in determining reliability of a test using CTT, there are different methods that can be employed such include test retest, parrallel form, scorer, split half Kuder-Richardson formulas 20 and 21 and Cronbach alpha methods. Despite the method used, CTT considers and provides only one source of error in a measurement at a time. For instance for test retest method, it assumes that the only source of error is occasion of testing, equivalent form method provides only one source of error in relation to forms of the test then internal consistency considers only the items. Shavelson and Webb (2005) asserted that the inability of CTT to separate the error score into different sources affected the generalization of the result obtained adversely. However, there are some measurement situations that involve the probability of the existence of multiple sources of error. In such cases, the application of CTT may not be feasible but may required any measurement theory in which more than one sources of error can be differentiated so that a technically weak test will not be used for decision making. This limitation of CTT paved way for the use of generalizability theory in estimating the reliability of a test.

GENERALIZABILITY THEORY (GT)

Generalizability theory is the statistical theory that uses factorial (random-effects) analysis of variance procedures to identify and estimate different sources of measurement error in an observed score that may in one way or the other influence the measurement of behaviour. GT identifies the different sources of both systematic and random variations, separate them and that of their interaction (Shavelson & Webb, 2005) Generalizability theory is a measurement theory that aimed at estimating the reliability of measurement obtained from different instruments or devices such as achievement test, rating scales and observation tools (Alkharusi, 2012).

GT assumes that data to be analysed must either be interval or ordinal in nature. Again that a student's observed score is made up of universe score and multiple sources of errors. Hence in GT, the reliability coefficient is expressed mathematically as;

$$\text{G-coefficient} = \frac{\text{Universe score variance}}{\text{Universe score + Individual source of error variance}} \quad \ldots \text{ equation 2}$$

So comparing question 1 and 2, it could be deduced that the true score in CTT was replaced with universe score in GT, while the undifferentiated (inseparable) error score in CTT was replaced with multiple sources of error score. So in GT, $O_s = T_s + E_{s1} + E_{s2} + E_{s3} + E_{s4}$ where O_s is the observed score, T_s is the true score E_{s1}, E_{s2}, E_{s3} are error scores from component 1, 2, 3 and so on.

In GT, the students or the testees are the object of measurement that is the person to be measured while the test score is a sample from a universe of admissible observations (Shavelson & Webb, 2005). So each student's observed score is broken-down into different components such as component for student, item, occasion and/or rater depending on the nature of the study. The students' component of the score is not a reflection of error but the systematic variations in students (individual difference among students. Then the other score components; item, occasion, rater and their interactions reflect sources of measurement error. This is an indication that GT assumes that in a measurement process, error may emanate from one or more of the following sources, the test items, testing occasions, test forms, the rater, and their interactions (such as PxI, P xO, PxIxO etc).

In other words, GT identifies and estimates the components of the individual observed score attributed to the student/examinee, the facets and their interactions. The facets are the characteristics of the testing conditions which represent the sources of variations such as the tests forms, test items, rater and occasion that exist in levels.

The levels of the facets are known as the condition. So "facets" and "condition" are the same as "factors" and "levels" respectively. Then the universe is the possible combination of the levels of the facets. It is the combination of the facets of observation that determine the condition to which the decision maker wish to generalise from a measurement to behaviour in the universe.

UNIVERSE OF GENERALISATION

In GT, the concept of reliability is applicable to either simple or complex universe depending on number of characteristics of testing conditions the decision maker intend to investigate (facets). Specifically, there are one-facet, two-facet, three-facet and more than three-facet universes.

One-facet universe is the universe of study where only a source of measurement error will be investigated. That is it is a design in which the universe of admissible observations and that of geeneralisation involve the same condition known as item facet. Item facet can either be denoted as "i" or "I" based on the reference to be made, if the reference is to be made to G-study it will be denoted as 'I' but if it is to made to D-study then item facet will be denoted as I while the object of measurement student/person is denoted as S or P respectively depending on the term used.

Brennan (2001) stated that there are two possible designs that could be applied in a G-study. They are the crossed and nested design that is PXI or the I:P design where P is the person (student or examinee)

i	is the item
x	is crossed with
:	is nested within

Based on the indexes, in the PXI design each student or examinee is tested on the same sample of test items. For I:P design each student is tested using different sample of test items. Each examinee is expected to have two scores from either the same test items administered on two different occasions or from the administration of two different forms of test items.

Two-facet universe is a study where the universe of admissible observations could be determined by the combination of two different facets such as items and occasion, where the universe of admissible observation are to be determine from all acceptable items that are administered at different points in time. A two-facet universe that is items x raters indicate that universe of admissible observations are to be determined using all acceptable items that are rated by different raters.

Sometimes, the complexities of a measure may not be determined by only two facets but more than two, in such cases a three or more faceted universe is required. So three-facet universe is a study in which the decision maker or test users tends to generalize or investigates the variability of test performance over three or more facets such as (items, occasion and test combined). Then the universe of observation will be determined from all possible items that can be given by all possible points in time by different test administrators.

Furthermore, generalizability theory is applied in two distinguished stage in terms of studies, generalizability study (G-study) and decision studies (D-study) to determine the dependability of scores obtained from measurement of behaviours. The G-studies provide the estimation of the generaliability coefficient of the variances from all possible facets while the D-studies help the test users to determine the coefficients among all possible interactions (Kane, 2002, Brennan, 2011).

Generalisabiliy theory also help to distinguished between two types of error variances associated with behavior measurement. These are the relative error variance and absolute error variances used to make relative and absolute decisions respectively. Relative decision entails the consistency of scores used in ranking students based on the differences in their performances in a given test. This error variance is associated with norm-reference interpretation of scores. So in all, relative error variance is the difference between a student's observed deviation score and his/her deviation from the universe score.

In contrast, absolute error variance is used to index the absolute level of an individual's performance in relation to the predetermined level without making reference to the performance of other students/individuals. This is mostly associated with criterion- reference interpretation.

Generalization theory also provides two reliability indexes terms generaliabiity coefficient and dependency coefficient. Generalizability coefficient is a measure of the estimate of the proportion of variance in a set of scores that are systematic for test designed for relative decision. Thus it is otherwise known as a norm-reference reliability or relative reliability that shows how accurate the generalisation of a person's observed score in relation to his or her universe score. It ranges from 0 to 1 where higher values indicate more dependable measures.

Dependability coefficient is a measure of the estimate o the proportion of variance in a set of scores for test designed for absolute decision, so it is sometimes termed as absolute reliability or criterion reference reliability where the cut-off score is set to the mean performance of the group. Like the generalizability coefficient, dependability coefficient (ф) has a ceiling of 1.00 while higher values represent greater precision. Generalizability and dependability coefficients are determined after the identification and estimation of the weights of various sources of error components.

Sequel to all these, generalizability theory primarily aimed at generalizing the scores from a specific group on a given measure to the universe of admissible observation and G studies, as well as the universe of G-studies and D-studies. Considering all the features of generalizability theory, it is deduced that it has the following advantages over the CTT in

1. It provides estimates for all the distinguished sources of measurement errors individually and in their combined effects using factorial ANOVA.
2. It provides detailed information concerning the generalizability studies and decision studies.
3. It enables the test developers and users to determine how many occasions, test forms, items and raters that are needed to promote dependable result for decision making (Yin & Shavelson, 2008).
4. GT recognizes that test users may be involved in two main different types of decisions based on the obtained scores. Thus, it distinguishes between relative error variance and absolute error variance leading to relative and absolute decision.
5. It also help to provide the estimate for each examinee's structural level of knowledge based on the examinee's performance in a test.

6. It provides estimate of reliability coefficients for test-retest, inter rater equivalent form and internal consistency.

Despite all the potential benefits of GT in measurement of behaviour, it is rarely used in reporting results of measures while CTT that does not have the ability to separate the multiple sources of measurement error but entangled them is widely used (Teker, Guler and Uyanik 2015). Baird and Black (2013) observed that public examination violates the assumption of item independence, normality of scores, and unidimensionality. Again that some of the public examinations are technically weak tests due to the establishment of their internal consistency, occasion-related factors and inter-rater reliabilities using, CTT.

In conclusion, Baird and Black (2013) stated that the use of CTT has made the field of educational measurement to be under-theorized so they suggested the use of a more flexible, powerful and better theory to take account of the educational context of public examinations which WASSCE chemistry essay test is one. Owing to this, the researchers were compelled to embark on the present study, which only focused on the two-facet universe where the a chemistry essay questions conducted by west African examination council in May/June 2018 was administered once on the SS3 students and their responses were rated by two different raters independently.

The rationale had been that, it is hoped that through the findings from the study a clear understanding on the limitation of the common method o establishing reliability will be made through the identification o the multiple sources of measurement error. Again, a test that its result will help to accurately determine the future technologists, medical doctor, engineers, and sound chemistry teachers may be developed. In another angle, the findings from this study may help to increase the confidence people have on the test result and as well promote transparency in the field of measurement by informing the test users the true ability of the students, which will in turn enhance productivity among future employee's of labour. This is because through the administration of a test with accurate dependability level, classification of students in their various areas of specialization will be effectively done. Finally through the findings from this study decision about the students and the subject will be substantially improved upon.

AIM AND OBJECTIVES OF THE STUDY

The aim of the study was to determine the reliability of WASSCE 2018 chemistry essay test using generalizability theory. Specifically the study sought to

1. Identify and estimate the magnitude of the variance component of chemistry essay test of 2018 May/June conducted by WASSCE.
2. Estimate the relative and absolute error variance, universe score, G-coefficient and D-coefficient of the chemistry essay questions conducted by WASSCE in May/June 2018.

EMPIRICAL STUDIES RELATED TO GENERALIZABILITY THEORY

In the past, there are various studies related to generalizability theory (GT) that were conducted. For instance, Solanor Flores and L_i (2006) conducted a study on the use of generalizability theory in the assessment of linguistic minorities among students. A two-facet, studentraters x items random study was conducted using three different sample sizes that were administered test on standard English, standard Haitian-crete and local chilect of Haitian-creole. Results from data analysis using urGENOVA, showed that in all the samples used across the different languages the largest percentage of error variance emanated from the students and item interaction. This was followed by that of the main effect, students, items and then the three-way interaction of students, raters and items. On the other hand, it was found that the variance component for raters and the interaction of students by raters contributed little or nothing to the total error variance.

Heitman, Kovaleski and Pugh (2009) used generalizability theory to estimate the reliability of ankle complex laxity measurement across different examiners and multiple trials. It was found that high measurement error was attached to the facets associated with raters than with trial for both anteropositerior and inversion aversion trials.

In 2013, semmetroth measured sources of variance on a special education teacher observation tool using generalizability theory by two facet partially rested design, in which occasions were nested within teachers and crossed with raters. It was reported that multiple sources of errors affected the levels of reliability of the special education. Specifically, it was found that the largest variance component emanated from the interaction effect of occasion nested within teachers and crossed with raters. This was followed persons, the interaction of person and raters and then raters.

The study conducted by Yelboga (2015) considered the estimation of the variance component of a proficiency examination in two different situations, cross pattern and mixed pattern using different programs such as GENOVA, EDuG, SPSS and SAS. It was reported that the percentage of variance component for all source obtained for all the programmes were consistent. That means in all the programmes multiple component of variances were obtained and the values were consistent. It was also specifically reported that the variance component for the three-way interaction of person, task and evaluate was the largest, followed by that of person and task, person, task, task and evaluator, person and rater and then lastly the evaluators.

Furthermore, Teker, Guler and Uyanik (2015) compared the effectiveness of spss and EduG in estimating the component of variance of a nine-item statistic test using three raters which gave rise to a two-facet design. The researchers used two designs, the fully crossed and nested random design, that is (P x I x R) and (P x I: r) respectively. It was found that the variance components estimated for the main effects, person, item and raters, two-way interaction effects for person and item (p1), person and raters (Pr) and item and raters (ir) as well as for the three-way interaction effects of person, item and raters (Pir) were approximately the same with both SPSS and Edu-G programmes. This trend was the same in both the crossed and nested random designs. In addition it was also reported that the largest variance component was attached to the main effect students. This is followed by that of the two-way interaction effects between students and item (S1) and thirdly by the three-way interaction effects of students, items and raters. On the whole the main effects of items and raters contributed little or nothing to the total error variance.

Nevertheless, Mushquash and O'Connor (2006) in their two-facet fully-crossed design study reported that the multiple sources such as items, (1) person and item (Pi), person and occasion (Po) and the three-way interaction effects of person, item and occasions (PiO) contributed to the variance in the undergraduate students scores in Rosenberg self-esteem scale. However, that the largest contribution to the total variance was that of three-way interaction effects of person, item and occasions (pio). This is followed by that of person, person by item interaction, item, person by occasions interactions and lastly by the items by occasions interaction. It was also reported that error variance in relation to the relative decision was a little below that of the absolute decision while the G-coefficient for relative decision was a little greater than that of the absolute decision.

An analysis of the previous related studies reviewed indicated that multiple sources of variances contributed to the error in a measure and that no indigenous study like the present study has been conducted. Again, none of the study was conducted in relation to chemistry achievement test. Thus, more so, CTT assumed that all observed scores are components of a true and an error score that are entangled. It is also worthy of note that with CTT, Gugiu et al (2012) asserted that it is difficult to obtain high reliable scores from essay test when rated by more than one raters. On these bases, it becomes very imperative to adopt other better approaches of determining reliability that will help to separate the various components of errors, minimize measurement errors and then increase the probability of getting high reliable scores. Precisely, the psycho-metricians had recently recommended the use of generalizability theory. This is because GT assumed that the error variance results from multiple sources and their determinations reveal the level of accuracy and dependability of the scores obtained from the measures. To crown it,Brennan (2001) stated that GT provides examination of the various sources of influence on score reliability within a single analysis.

METHODS

The study adopted a two-facet fully crossed random design (S x I x R). This is because all the students used for the study responded to all the items and all their responses were doubled scored using two random, blinded and independent scorers Shavelson and Webb (1991) stated that in a crossed design every person responds to the same set of items which will be rated by all the raters independently. So in this study which is denoted as (S x I x R), S is the students which denote the object of measurement and not facet. 'I' denotes the five items administered to the students while "R" is the two raters. Thus the items and raters are the two facets. The crossed (X) symbol denote that all the items will be responded by all the students and all the raters must also rate all the items independently.

A sample of 74 senior secondary three chemistry students, in Obio/Akpor Local Government Area of Rivers State, Nigeria was used for the study. They were obtained using two-stage sampling method where at the first stage, simple random technique by balloting was used to select six senior secondary schools in the area.

At the second stage, non-proportionate stratified random sampling and accidental/ convenience sampling techniques were used to select 15 senior secondary three students from each school irrespective of the size of chemistry students in each of the six chosen schools. It is worthy of note that accidental/convenience sampling technique was used because the 15 senior secondary three chemistry students were selected based on their availability and willingness to respond to the items in the test. On the whole 90 Senior Secondary three chemistry students were obtained. However after test administration, during scoring and collation it was observed that 16 students did not answer all the items, so they were removed and 74 senior secondary three chemistry students were then used for the study. This sample of 74 senior secondary three chemistry students is adequate for the study. This is because Alilgan (2013) recommended a sample of 50 to 300 to be adequate for unbiased adequate estimation of the coefficient and phi-coefficients.

Furthermore, for data collection, an adapted instrument tagged WASSCE chemistry paper 2 question conducted by West African Examination Council for school candidates in May/ June 2018 was used. It is made up of 5 essay (open ended) questions with two sections (A and B). section A is made up of only one question for all candidates in all the countries. Then the section B, which is country based has four questions to answer only three. However, on the basis of two-facet model of generalizability theory, the students were asked to answer all the four questions in that section for Nigeria candidates plus the one question in section A given a total of 5 questions.

The face and content validities as well as the reliability of the instrument were not estimated on the basis that examinations conducted by West African examination council (WAEC) are a standardised test. During the administration of the instrument a direct-delivery approach was employed. The copies of the instruments were administered by the researchers and the assistant of the chemistry teacher of each school. The rules and instruction governing the examination were strictly followed except that time allowed which supposed to be 2 hours was changed to 2½ hours. This is based on the shift of answering only 4 questions to answering all the 5 questions in compliance to the two-facet fully crossed design.

To maintain standard, the responses of the students were rated using two independent raters who are WASSCE chemistry paper 2 examiners in Port Harcourt Rivers State, Nigeria. These raters utilized the 2018 chemistry II marking scheme as a guide to their scoring/rating:

Data obtained from their scoring were subjected to Scientific Package for Social Science (SPSS) to run a student x item x rater random effect analysis of variance by univariate model. This was used to partition the total variability in the data set into its separate sources of variations for the object of measurement (students), components for the main effects sources such as students, items and raters and their interactions. This is because the score given to each student by each rater on each item is conceived to be the deviation from the grand mean over all students, raters and items where the degree of deviation was determined by students (object of measurement) effect in the form of universe score, items' and raters' effect.

In addition to the three main effects each facet, item and raters' interaction with the object of measurement (students) in a two-way interaction SI and SR respectively as well as the two-way interaction effect between the two facets items and raters (IR) were estimated. There was also a three way interaction effect among the object of measurement (S) items and raters (SIR) and that of error component.

To obtain the variance estimate for each component the data were subjected to factorial analysis of variance by variance component procedure. This is as suggested by vispoel, Morris and Klinc (2018). They stated that variance components for G-theory analyses can be computed using the VARCOMP procedure in SPSS via univariate model. This can also be done alternatively, by setting the observed mean squares from the ANOVA to the expected mean squares equations. After obtaining the variance components estimates, the percentage of the total variance for each variance estimate was determined by dividing each variance estimate by the total variance and multiplied by 100.

In determining the estimate of the error variances for relative and absolute conditions the following formula as stated in Vispoel et al (2018) were used.

Relative error variances = $\frac{\sigma^2 pr}{n_r} + \frac{\sigma^2 pi}{n_i} + \frac{\sigma^2 pr}{n_r\, n_i}$

Absolute error variances (σ absolute) =

$$\frac{\sigma^2 r}{n_r} + \frac{\sigma^2 i}{ni} + \frac{\sigma^2 pi}{ni} + \frac{\sigma^2 pr}{n_r} + \frac{\sigma^2 ri}{ni\, n_r} + \frac{\sigma^2 SIR}{ni\ n_r}$$

Where $\frac{\sigma^2 pr}{n_r}$ = transient error variance

$\frac{\sigma^2 pi}{ni}$ = specific-factor variance

$\sigma^2 pri$ = Random-response error variance

RESULTS

After data analysis the results obtained for research question 1 and 2 are presented in tables 1 and 2 respectively.

Table 1: Estimated variance components and percentage of score variation for WASSCE on chemistry essay question 2018.

Source of variation	Sum of squares	df	Mean square	Variance component	% variance component
Student(s)	2828.81	73	39.75	0.481	2.36
Item (1)	901.72	4	225.43	1.285	6.31
Rater (R)	19.46	1	19.46	0.039	0.192
S x I	9855.28	292	33.75	15.27	75.04
S x R	249.14	73	3.41	0.0378	0.186
I x R	18.89	4	4.72	0.020	0.098
S x I x R	94.51	292	3.221	3.221	15.83
Error	0.000	0	0.000	0.000	0.000
Total	14813.81	739	-	20.35	

The results in table 1 revealed that the estimated variance for the object of measurement (students) is 0.481, which accounted for 2.36% of the total variance in the chemistry essay questions scores. For items, its variance component estimate is 1.285 which is equivalent to 6.31% of the total variance while that of the raters is 0.039, which constituted 0.19% of the total variance.

Moreso, in table 1, it is also revealed that variance components for all the two-way interactions between each of the facets and object of measurement were also obtained. For instance, it is shown that for interaction between students and items (SI) an estimate value of 15.27, which represent 75% of the total variance in the students score in the WASSCE 2018 chemistry essay questions was obtained. For the interaction between students and raters (SR), an estimate value of 0.0378, which accounted for 0.186% of the total variance, was obtained. Then for the interaction between the two-facets, item and raters (IR) an estimate value of 0.020 equivalents to 0.098% of the total variance was obtained.

Finally, in the same table1, it is shown that the variance component for the three-way interaction effects among students, items and raters (SIR) is 3.221, which is equivalent to 15.83% of the total variance in the students' score in chemistry essay question conducted by WAEC 2018.

Table 2: Estimated generalizability coefficients relative and absolute decisions

Relative error variance	Absolute error variance	Universe score	G-coefficient	D-coefficient
0.124	0.116	10.02	0.795	0.806

Table 2 shows that values estimated for relative error variance and absolute error variance are 0.124 and 0.116 respectively while that of the universe score is 10.02. Furthermore, a critical observation on table 2 revealed that estimated values obtained for G-coefficient (P^2) and D-coefficient (index of dependability) (θ) are 0.795 and 0.806 respectively.

DISCUSSION OF FINDINGS

Generalizability theory is used to determine the reliability of instrument when multiple sources of variations contribute to measurement error. That means generalizability theory helps to disentangle the multiple sources of error in a given measure. In this study, the result obtained after data analysis, revealed that multiple sources contributed to the measurement error in the chemistry essay questions conducted by West

African examination council in 2018. The sources of measurement error include the items, raters, the interaction effects of SI, SR, IR and SIR. However a critical evaluation of their variance components revealed that the largest variance component was obtained from the two-way interaction effects between the object of measurement(s) and the item (SI). This implies that the relative standing of the students differ across items. In other words the students did not score high in all the items leading to great variations in their performances across the items. This finding suggests that difficulty levels of the items greatly differ and that the students also differ in their ability levels.

The next largest variance component emanated from the three-way interaction effects among the main effects student, items and raters (SIR). The estimated value obtained for this source of variation (SIR) indicates that students relative standing differ across the items and the rating of the students in all the items by the raters also differ to a great extent. This implies that the different raters rated the students differently across items. The raters' levels of agreement in their ratings differ very well across the students. However, this finding to some extent supported that of Yelboga (2015) since both study reported that SIR contributed a large measurement error to the total variance.

It was also found that the variance component for item (1) was reasonably high. This means that part of the measurement error came from the item, which implies that the mean performance of the students differ from item to item. This variation in the mean performance of the students across the items may be attributed to differential difficulty levels across the items. Again, it could be traceable to the relative standing of the students which vary from one item to another.

Furthermore, from the study, it was found that the variance component for students was some-what large. This implies that the universe score among the students vary from one person to another. So since students represent the object of measurement and not error, it is then deduced that the level of variance component obtain from students represent the systematic individual differences in chemistry to a reasonable extent.

Again from the study it was also found that other sources of measurement error in chemistry essay questions conducted by WAEC 2018 such as raters, interaction effects between students and raters (SR) and that of items and raters (SR) have very small variance components. Thus they contributed very small to the measurement error in the chemistry essay question. However, the small variance components for these sources may suggest that inaccuracies in generating the scores of the students are very small. This may be due to the use of similar marking scheme by the raters. Hence, there is a high level of agreement between the raters in the relative standing of the students in a given item. In other words the two raters were somewhat comparable in their ratings of the students' performance in the chemistry essay questions. This finding is not in line with that of Heitman et al (2009). They reported that a large measurement error was attached to raters. The two findings also differ on the bases of instruments and respondents that both study considered.

It was also found that the unmeasured error component contributed little or nothing to the total variance. This may suggest that the sources of measurement errors in the chemistry essay questions as separated by G-theory analysis have been indentified and little or no other sources yet to be identified based on 2-facet study design.

Nevertheless, from the study it was found that little inconsistency was observed in the ranking of the students by the raters based on their performances in the chemistry essay questions. Considering the absolute error variance estimate, it was revealed that the students observed scores did not deviate so much from the cut-off point which is the universe score.

With respect to the generalizability coefficients obtained, it is obvious that the students' relative standing can be differentiated with high degree of accuracy despite the random fluctuation of the conditions of the measurement. This is because the estimated proportion of the students observed score variance due to their universe score variance is quite high. Finally, the dependability coefficient obtained implies that a higher precision of accuracy was recorded in the performance level of students in relation to the predetermined cut-off point. Hence, the result obtained yielded higher precision level of accuracy in both relative and absolute decision making concerning the students. This is because the variance components corresponding to the main effects items and persons/students are quite high.

RECOMMENDATIONS

On the basis of the findings the following were recommended.

1) The use GT in determining reliability/dependability of measuring instruments should be emphasized by the examining bodies and institutions of learning.

This is based on its ability to disentangle the multiple sources of error.

2) Researchers and test developers should endeavor to report the reliability of scores from measurement scale without underestimating the contributions of multiple sources of error variance in a set.

3) Information obtained from WASSCE paper 2 conducted in May/June 2018 should be depend on.

4) The use of marking scheme should be encouraged as it helps to reduce the measurement error that would emanate from raters.

5) Items with comparable difficulty levels should be used to make up a given test.

CONCLUSION

Owing to the findings from this study it is conceivable that generalizability theory under the two-facet model is one of the most sophisticated methods of determining reliability of scores. Again it had been demonstrated that 2-facet model of generalizability can be used to estimate and integrate reliability. It is also in limelight that chemistry, essay question conducted by West African Examination Council in 2018 is reliable, and hence scores obtained from it are dependable. It is also confirmed that generalizability theory is a method of estimating reliability that can disentangle the various error components in a measurement.

REFERENCES

1. *Alkharusi, H. (2012). Generalizability theory: On analysis of variance approach to measurement problems in Education Assessment. Journal of Studies in Education, 2(1) 184-196.*
2. *Anastasi, A. & Urbina, S. (2006). Psychological testing. New Delhi: Pearson Education incorporation.*
3. *Anatol, T. & Hariharan, S. (2009) Reliability of the evolution of students' answers to essay type questions. West Indian Medical Journal, 58 (1) 13-16.*
4. *Ary, D. Jacobs, L.C. & Razavieh, A. (2002). Introduction to research in education (6th Ed.) Australia: Wedsworh Thomson Learning.*
5. *Atilgan, H. (2013). Sample size for estimation of g and phi coefficient in generalizability theory. Eurasian Journal of Educational Research, 51, 215-228.*
6. *Baird, J. & Black, P. (2015). Test theories, educational priorities and reliability of public examinations in England: Journal of Research Papers in Education 28(1) 5-21.*
7. *Brennan, R.L. (2003). Coefficients and indices in generalizability theory center for advanced studies in measurement and assessment. CASMA Research Report.*
8. *Brennan, R.L. (2011). Generalizability theory and classical test theory. Applied Measurement in Education 24, 1-21.*
9. *Brown, G.T. (2010). The validity of Examination essay in Higher Education: issues and responses. Higher Education Quarterly, 6424(3) 276-291*
10. *Elliot, S.N. Kratochwil, T.R, Cook, J.L. & Travers, J.F. (2000). Educational psychology: Effective teaching, effective learning. Boston McGraw-Hill.*
11. *Gugiu, M.R., Gugiu, P.C. & Baldus, R. (2012). Utilizing generalizability theory to investigate the reliability of grades assigned to undergraduate research papers. Journal of multi-Disciplinary Evaluation 19, 26-40.*
12. *Heitman, R.J., Kovaleski, J., Pugh, S.F. (2009). Application of generalizability theory in estimating the reliability of Ankle complexity measurement. Journal of Athletic training 44(1) 48-52.*
13. *Kane, M. (2002). Inferences about variance component and reliability-generalizability coefficients in the absence of random sampling. Journal of Educational Measurement, 39 (2), 165-181.*
14. *Mushquash, C. & O'Connor, B. P. (2006). SPSS and SAS progress for generalizability theory analysis. Behaviour Research Methods 38(3), 542-547*
15. *Onukwu, G.I.N. (2002). Fundamentals of educational measurement and evaluation. Owerri: Capes Publishers.*
16. *Orluwene, G.W. (2012). Fundamentals of testing and non-testing tools in educational psychology. Port Harcourt: Herey Publication Coy.*
17. *Rust, J. (2007). Discussion piece: The psychometric principles of assessment. Research Matters: A Cambridge Assessment publication, 3, 25-27.*
18. *Schuwirth, L.W. & Vieuten, C.P. (2011). General overview of the theories used in assessment. Medical Teacher, 33(10), 783-797.*
19. *Semmelroth, C.L. (2013). Using generlisability theory to measure sources of variance on a special education teacher observation tool. Boise State University: Unpublished Doctor of Education Thesis*
20. *Shavelson, R. J. & Webb, N.M. (2005). Generalizability theory: An overview. In B. S. Everith & D. C. Howell (Eds.), Encyclopedia of Statistics in Behavioral Science (pp.717-719). Chichester: John Wiley & Sons*
21. *Solano-Flores, G. & Min, Li (2006). The use of generalisabiliy (g) theory in the testing of linguistic*

minorities. Educational Measurement: Issues and practice 25(1)

22. *Teker, G.T., Guler, N. & Uyanik, G.K. (2015). Comparing the effectiveness of SPSS and EduG using different designs for generalizability theory. Education Sciences: Theory and Practice 15(3), 635-645.*
23. *Vispoel, W.P, Morris, C.A. & Kiline, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. Psychological Methods 23 (1) 1-26.*
24. *Yin, U. & Stiavelson, R.J. (2008). Application of generalizability theory to concept map assessment. Research Applied Measurement in Education 21:273-291.*