Application of generalizability theory in estimation of variance components in National Examination Council essay questions in English Language

[1]Ogunka, Richard Iheanyichukwu [2]Dr. Goodness Orluwene

Abstract

Application of generalizability theory in estimation of variance components in National Examination Council Questions in English language. Design is two facet fully crossed G – study and D – study. The population 723 senior secondary schools in Ikwerre Local Government Area. Simple random sampling technique via balloting to draw six schools out of fourteen senior secondary schools. Cluster sampling technique to sample 153 students. Instrument is 2018 November/ December National Examination Council English Language Essay Questions. Data obtained analysed using computer software SPSS through General Linear Model via variance components MINQUE method. Result the largest contribution to error is person by item by rater (σ^2pir) (5.259) with percentage variance of 47: 221 and person by item (σ^2pi) (4.253) percentage variance of 38.17; person by rater (σ^2pr) (0.415) percentage variance of 3.72, person (σ^2p) (0.398) percentage variance 3.57%,rater (σ^2r) (0.80) percentage variance 7.18, item (σ^2i) (0.024) percentage variance 0.21, item by rater (σ^2ir) (1-0.008) percentage variance -0.071. The universe score 5.5368, Relative error variance 5.1146, Absolute Error Variance 0.4222, G- study coefficient 0.5198 and Index of dependability 0.929 obtained. Recommendation, generalizability theory should be used in psychometric properties estimate to generate reliable test items.*

Key word Essay, Variance, Estimation, Generalizability theory.

Introduction

English Language is a core subject in Primary and Secondary Schools in Nigeria Educational Environment. It is language of communications in both school and non-school settings, English Language has to be learnt in schools because of the role it plays in the educational system and the society. The teaching and learning of English Language must not be taken for granted, it should be carefully and thoroughly studied. Adegbile (2002; 2003) clearly stated that success at every level of educational system depends largely on competence in English. Certificate awarded to candidates that pass the examination set by National Examination Council could be used to obtain employment and again the minimal requirement for admission into any tertiary institution of higher learning in Nigeria points at credit grade in English Language. The senior secondary school certificate examinations conducted by National Examination Council have shown that many students fail ridiculously in English Language papers. Issues that bother the teachers, the public, parents and other stakeholders is mass failure rate which they blame on the National Examination Council as it concern construction, validation, and administration of examination conducted by them that lack adequate psychometric analysis of its items; this research is needful based on public complain and perception on the recorded mass failure, hence development of a measuring instrument that accurately measures a particular characteristics of the examinee in the best perfect manner without any unsystematic or systematic error both in the instrument and characteristics under investigation is bound to pose problems to the examinee. It is only when the instrument measures in perfect manner that such unquestionable quantitative descriptions of the examinee in terms of the exact extent to which it possess and demonstrate the trait can be adjudge for the best in relation to decisions to be subsequently taken for reliable oral, essay and objective item in English Language. Hence this investigation seek to validate essay question.

Essay Questions is one of the assessment tools in testing or assessing student's achievement in any

given instruction. Essay tests measures the higher levels of cognitive domain which provide critical thinking and originality in students. Linn, Miller and Gronlund (2005) stated that essay items provide the freedom of response that is needed to adequately assess the ability of students to formulate problems, organise, integrate and evaluate ideas and information and thus, apply knowledge and skills. Orluwene (2012) defined Essay test as the type of test item which allows the students the freedom to supply their own responses, rather than select the correct answer. She further stated essay tests are presented with a narrative or question form, and the students are required to compose a reply which presents a complete response in at least one sentence; since the item allows for task with larger scopes, by requiring students to organize and integrate information, interpret information, give arguments, give explanations, evaluate the merit and demerit of ideas, and conduct other types of reasoning that tap complex thinking. In this study, the researcher will make efforts to avoid the use of options in English Language essay questions because it is the assessment of levels of all the sampled population to estimating variance components in National Examination Council English Language Essay Questions.

Generalizability (G) theory is a statistical theory about the dependability of behavioural measurement, Cronbach, Gleser, Nanda, and Rajaratnam (1972) states the notion of dependability as follows: The score (on a test or other measures) on which the decision is to be based is only one many scores that might serve the same purpose. The decision maker is almost never interested in the response given to the particular stimulus objects or questions, to the particular tester, at the particular moment of testing. Some, at least, of these conditions of measurement could be altered without making the score any less acceptable to the decision maker… The ideal data on which to base the decision would be something like Person's mean score over all acceptable observations. (p.15). Kin and Wilson(2009) refers dependability of behavioural measures as the accuracy of generalizing from a person's observed score on a measure, or a test to the score that the person who have received average 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. The unsystematic error or fluctuation in the individual's scores over several repeated testing's mean that in the behavioural sciences, one cannot completely depend on the single score obtained by each student on an attribute that was measured once. According to Shavelson and Webb(1991) dependability refers to the accuracy of generalizing from a person's observed score on a test or other measure (behaviour observation, opinion survey) to the average score that person would have received under all the possible conditions that the test user would be equally willing to accept. This notion of dependability is the assumption that the person's knowledge, attitude, skill, or other measured attribute is in a steady state; it is assume that any differences among scores earned by an 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 (2012) indicated that, in the measurement of complex traits imperfect instruments are used so that the score observed for each person almost always differs from 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. To satisfactorily control the whole issue raise above on the problem of incomplete certainty in and dependence on a score obtained by an individual from a single administration of a test for accurate quantitative description of the person with respect to a given psycho- social construct, the researcher is required to empirically establish the reliability of the measuring instrument which he or she develops. The comparism of dependability of reliability in generalizability theory and classical test theory to determining standard error measurement varies. Atilla(2012) asserted that the use of classical test theory approaches to determining score reliability, however, are not capable of

identifying and untangling this profusion of error which classical reliability was not conceptualized to do since it account for only one source of error at a time. Similarly, Ikeh and Madu cited Tavako and Brennan (2013) states Classical Test Theory (CTT), assume the student's true score is the sum of the student's observed score and a single undifferentiated error term. Kpolovie (2010) asserted classical theory as reliability embedded in the true- score and error- score model defines reliability as the coefficient that predictable proportion of variance in observed scores from the true scores. Generalizability Theory liberalizes classical theory by employing ANOVA methods that allows an investigator to untangle multiple sources of error that contribute to the undifferentiated E in classical test theory. But is important to state that GENOVA, SAS, SPSS and Edu-G program are computer software used for statistical analysis, data mining and predictive analysis. In this study SPSS computer program via General Linear Model on Variance Components (MINQUE) method was fully adopted for the estimation of variance components for generalizability theory. Generalizability Theory is a statistical theory for estimating the reliability of behavioural measurements which offers researchers an opportunity to asses comprehensively sources of measurement error (variance components). G- Theory concerns itself about the relative and absolute dependability of behavioural measures. The empirical works that support this investigation includes, Atilla (2015) estimation of generalizability coefficient: application with different programs, Ikeh and Madu (2018) applications of generalizability theory in estimating multiple sources of variation in economics essay test, Heitman, Kovaleski and Pugh (2009) Application of generalizability theory in estimating the reliability of Ankle-Complex Laxity Measurement and Preuss (2003) Using generalizability theory to Develop Clinical Assessment Protocols. Items – crossed – with raters in the universe of admissible observations. This means that any one of the Ni items might be rated by one of the Nrraters. If the population is crossed with this universe of admissible observation, then the corresponding G-study design is p X i X r, in which the responses of Np persons by Ni items are each evaluated by Nrraters (Brennan 2001). Therefore, estimation of variance components for this fully crossed design can be used to estimate results for any possible two facet design. The theory of generalizability focuses on the magnitude of sampling out errors due to person, item, rater and occasions among others, then their interactions which provides estimates of the magnitude of measurement error in the variance components as well provide a summary dependability coefficient reflecting the generalizing a sample score or profile to the much larger domain of interest (Shavelson, Baxter and Gao, 1993).

This study seems to thoroughly investigate person by item by rater variance of measurement error in the context of cognitive domain based test in National Examination Council Essay questions in English Language. Johnson, Dulanay and Banks (2002) asserted measurement error is a situation in which students true ability is either underestimated or overestimated. Hence, the need of estimating measurement error is inevitable because the inconsistencies that exist in measurements instruments is enormous especially variance components.

Aim

The aim of this study is to estimate the variance components of persons by items by raters and scores dependability on National Examination Council Essay Questions in English Language with the application of generalizability theory.

Research Question:

1. What are the relative contribution of person, i t em, rater and their interactions?
2. What is the dependability coefficient reliability of the instrument?

Methods

The design is instrumentation with two – facet fully crossed G – Study and D- Study. D- Study generalizability theory uses the information obtained from G-Study to determine the measurement procedure in minimizing undesirable variance and maximize dependability reliability.

The population are made up of all Senior

Secondary School Three Studentsin Ikwerre Local Government Area of Rivers State, Nigeria. Simple random sampling technique via balloting used to draw six schools out of fourteen senior secondary schools. Cluster sampling technique used to sample 153 SS3 Students from the population of 723 SS3 Students and was administered the National Examination Council Essay Question (continuous writing section) in English Language.

Instrument for data collection – 2018 November/December National Examination Council Essay Question in English Language mostly continuous writing which contains four items (Question), each item is rated on 10 points marking scheme which was adapted by the raters.

The data collected was analysed using a computer software SPSS through General linear model under Variance components MINQUE method.

Results

Estimate of Variances component on Generalized Linear model under variance components (MINQUE METHOD)

Source	Type 1 sum of variances	Df	Mean. Square	Variance component	% variance
Person	2829.114	152	18.613	0.398	3.57
Item	159.800	3	19.933	0.024	0.21
Rater	54.805	1	54.805	0.80	7.18
Person * Item	6277.075	456	13.766	4.253	38.17
Item * Rater	11.937	3	3.979	-0.008*	-0.071
Person * Rater	1051.820	152	6.920	0.415	3.72
Person * Item* Rater	2397.938	456	5.259	5.259	47.221
Error (Residual)	000	0		.000	
Total	12682.489	1223		11.141	100

components (MINQUE METHOD)

Table 1: The result presented the estimated variance components and its interactions. The result shows that the largest contribution to measurement error is person by item by rater (∂^2pir) (5.259) with percentage variance of 47: 221 and it is followed by person by item (∂^2pi) (4.253) percentage variance of 38.17; person by rater (∂^2pr) (0.415) percentage variance of 3.72, person (∂^2p) (0.398) percentage variance 3.57%, rater (∂^2r) (0.80) percentage variance 7.18, item (∂^2i) (0.024) percentage variance 0.21, item byrater (∂^2ir) (-0.008*) percentage variance -0.071.The estimation of various component interacted very high among person, item and rater(∂^2pir).

Estimation of G and D reliability coefficient

Relative Error variance	Absolute Error variance	Universe score	G-Study Coefficient	Index of Dependability
5.1146	0.4222	5.5368	0.5198	0.929

Table 2: The result shows estimation of G and D study reliability coefficient where.
The universe score 5.5368, Relative error variance 5.1146, Absolute Error Variance 0.4222, G- study, coefficient 0.5198 and Index of dependability 0.929, were obtained from the scores.

Discussion
The result revealed that the largest contribution to measurement error from the score obtained is on the person, item and rater ∂^2pir (5.259, 47.221%) the result indicated that a proportion of the variance was due to the interaction of person by item by rater (∂^2pir).However, this large variance component observed is not only in relation to persons, items and raters but also to undifferentiated error. The second largest source of variance is person by item ∂^2pi (4.253, 38.17%) Third, person by rater ∂2pr (0.415, 3.72%) relative to variation due to person ∂2p (0.398, 3. 57%) followed by rater ∂2r (0.80, 7.18%) then item ∂2i (0.024, 0.21%) The item by rater ∂2ir variance component has (-0.008*,-0.071%) indicating that error due to item by rater yielded negative estimate (-0.008), as a result of the degree of freedom for the residual (Error) amounted to zero. However the concern with variance component estimation is when a negative estimate arises because of sampling error or model misspecification, the possible solution is to set negative estimate to zero but use of negative estimates in expected mean square equations for other components (Brennan 2001).

Interestingly, the estimation of the G and D study of generalizability theory was achieved by using the universe score.A generalizability coefficient is the ratio of universe score variance to itself plus relative error variance, and Index of dependability is the ratio of universe score variance to itself plus absolute error variance (Brennan 2001).The universe score is 5.5368 on itself plus relative error variance estimate G. study coefficient 0.5198. Again the universe score of 5.5368 on itself plus Absolute Error variance estimate Index of dependability 0.929. Therefore, the reliability coefficient of the instrument is 0.93, which is high and adequate for certificate examination.

Conclusion
Based on the results of the generalizability analysis, the largest contribution of variance components obtained is from person by item by rater, followed by person by item, rater, person by rater, person, item. While item by rater indicated a negative estimate as a result of the degree of freedom for the residual which amounted to zero. Hitherto, the Index of dependability is high and reliable as much the instrument is good to use for certificate examination.

Recommendation
This study is basically to reduce the influence variance components that arise from scores by students in any examination. Generalizability Theory/analysis should be used to subject students' scores for psychometric properties estimate in order to generate reliable test item for examination bodies

Reference

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

Atilla, Y (2015) Estimation of generalizability coefficient: an application with different program.

Adegbile, J. A (2002) 'The Relative Effects of two model of Advance organiser onp e r f o r m a n c e i n R e a d i n g Comprehension '. In Journal of p e d a g o g y a n d E d u c a t i o n a l Development, Rivers State College of Education, Port Harcourt.Vol.9, No 2.pp, 22-36.

Brennan, R.L (2001) Generalizability Theory: Statistics for Social Science and Public

Policy. Springer-Verlag Berlin Heidlberg. New York.

Brennan, R.L. (2001). An essay on the history and future reliability from the perspective of

Replications. Journal of Educational Measurement, 38, 295 – 317.

Cronbach, L.J, Gleser, G.C., Nanda, H. &Rajaratnam, H. (1972). The

dependability of behavioural Measurements. New York: Wiley.

Heitman, R.J; Kovaleski, J.E & Pugh, S.F. (2009). Application of generalizability theory in estimating the Reliability of Ankle- Complex Laxity Measurement. Journal of Athletical Teaching. https://www.ncbi,nih.gov

Ikeh, E.F. &Madu, B.C. (2018) Application of generalizability theory in estimating multiple sources of variation in economics essay test. A paper presented at National Annual Conference (Abuja 2018). Association of Educational Researcher and Evaluators of Nigeria.

Johnson, S. Dulanay, C. & Bank, K. (2000). Measurement error. Retrieved from http:/ www. wcpss. net/ evaluation research/reports/2000/mment_error.pdf

Kim, S. C & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment Using generalizability theory. Journal of Allied Measurement, Edition USA Pearson teaching 9th Prentice Hall.

10(4) 408 – 423.Kpolovie, P. J (2010) Advanced Research Methods, Owerri Spring field Publisher ltd.

Linn, R. L., Miller, M. D. & Gronlund, N. E. (2005). Measurement and Assessment in

Shavelson, R. J. & Webb, N. M. (2005). Generalizability Theory.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling Variability of Performance

Assessments. Journal of Educational Measurement, 3(30), 215 – 232.R e t r i e v e d f r o m http://links.jstor.org/sici?sici=0022-0655%28199323%2930%3A3%3C215%3ASVOPA%3E2.0.C0%3B2-M

Preuss, R.A. (2003). Using generalizabilitytheory to Develop Clinical Assessment Protocols. Physical Therapy, vol. 93issue 4, pp 562-569