UNIJERPS
Unizik Journal of Educational Research and Policy Studies VOL.17 (3);
January-June, 2024 https://unijerps.org
2023 Impact Factor:* 6.834;** 2022 ***Impact Factor*:** 5.156; ***2021 Impact Factor: 5.641;
2020 Impact Factor: 5.355. *Indexing and Impact Factor URL: http://sjifactor.com/passport.php?id=21363*

ESTIMATING MEASUREMENT ERROR OF SENIOR SECONDARY SCHOOL ECONOMICS EXAMINATIONS USING GENERALIZABILITY THEORY IN IMO STATE

Chukwuemeka B. Ikeanumba
Phone: +2347039370884 **Email:** cb.ikeanumba@stu.unizik.edu.ng
Department of Educational Foundations, Faculty of Education
Nnamdi Azikiwe University, Awka

Nkechi P.M. Esomonu
Phone: +2347026422569 **Email:** n.agu@unizik.edu.ng@gmail.com
Department of Educational Foundations, Faculty of Education
Nnamdi Azikiwe University, Awka
&
Kelechi L. Anikpe
Phone: +2348034018582 **Email:** kl.anikpe@stu.unizik.edu.ng
Department of Educational Foundations, Faculty of Education
Nnamdi Azikiwe University, Awka

ABSTRACT

Scores obtained by the students (object of measurement) in examinations are affected by multiple sources of error (facets) and these scores, are used in taking relative and absolute decisions about students. With poor performance of students in Economics, there is need to estimate measurement error and dependability of the scores to find the extent of the contributions of these sources of error in examination scores. This study estimated measurement error of the 2020 SSCE Economics essay examinations using generalizability theory in Okigwe Education Zone of Imo State, Nigeria. To guide the study, two research questions were answered and one hypothesis tested at 0.05 level of significance. A random effect two-facet fully crossed s x i x r design for generalizability (G) and decision (D) studies was adopted for the study. The population of the study comprised 6,799 Senior Secondary Three (SS3) students from the 137 public secondary schools in the Okigwe Education Zone for the 2023/2024 academic session. The sample for this study comprised 680 students representing 10% of the total number of SS3 students and 68 schools representing 50% of the total population of schools in the zone. Paper 2 sections A and B of the 2020 WAEC Economics essay paper and 2020 and 2021 WAEC Economics Marking Guide were adopted for this study. The instruments were already validated and standardized by WAEC. Data collected and analyzed for research questions using the software EduG version 6.0-e based on analysis of variance (ANOVA) and generalizability. Hypotheses were tested using 95% confidence interval using standard error of measurement (SEM) and the variance components at 0.05 level of significance. The findings of the study revealed among others that, the students' facet (σ^2_S) made the highest contribution to measurement error in examination scores followed by the residual (σ^2_{SIR}), that is, the interactions

of students x items x raters; and the interaction between students x items (σ^2_{SI}) in the 2020 SSCE economics examination. Based on the findings, it was recommended among others that, Generalizability analysis should be carried out by test developers, psychometricians, policymakers in assessment in the estimation of reliability to identify multiple sources of error and to reduce or eliminate measurement error, hence maximize reliability of candidates' scores.

Keywords: Generalizability theory, instrumentation facets, and students

Introduction

Measurement is the process of assigning numerical values to describe features or characteristics of objects, persons, or certain events in a systematic manner. It involves assigning figures, numerical quantities, or scores to variables or traits of interest. Measurement spreads through almost every aspect of modern society because it is carried out by every individual both learned and unlearned. A great variety of things about the individual which includes achievement, aptitude, intelligence, height, and weight are measured by various people like teachers, doctors, and others, and are done regularly. Obtaining value from measurements seems quite simple, but a major problem with many kinds of measurement is that, there is often no basis to assume that the numerical values obtained, accurately and truthfully represent the trait of interest. The results of measurements can have profound influence on an individual's life, hence, it is important to understand how these scores are derived and the accuracy of the information they contain.

Measurement plays an important role in Education, Sciences, Engineering, and many other fields, as well as, in everyday activities. It involves assigning numbers according to a specific rule such as awarding points to various aspects of an achievement test. In education, traits like aptitude, opinion, scholastic ability, or achievement can be measured using different instruments like tests, questionnaires, interviews, and others. These instruments have various purposes which include description, prediction, assessment of individual differences, objective evaluation, domain estimation, mastery decisions, and diagnosis. Another important usage of educational measurement is to classify examinees, because errors in classification may lead to taking wrong decisions. In line with the assertion of Omoroguiwa (2019), educational measurement is the science and practice of obtaining information about characteristics of students, such as their knowledge, skills, and abilities, among others which will be used for educational purposes. The most common instrument in obtaining this information at the classroom level is achievement test (Esomonu & Ikeanumba, 2021).

Achievement test is a test used to determine the cognitive ability of students. It is also a test designed to measure the outcome level of accomplishment in a specified programme of instruction in a subject area or occupation that a student had undertaken. The belief is that a student who scores high

on the test has high ability level than a student who scores low. In the same vein, if two students' scores are similar, they are assumed to have the same level of ability. No test, however, well-designed can measure a student's true ability without error because numerous factors interfere with the ability to measure it accurately and precisely. These factors otherwise referred to as sources of measurement error or random error, include questions, examiners, invigilators, and students' guessing among others. This means that every score obtained by students has some amount of error in it. If assessment instruments or evaluation tools are faulty or error-prone, they yield invalid results which lead to inappropriate decisions about student educational outcomes (Adeosun, 2020). Thus, the goal of educators and psychometricians is to design a test that has as little or no measurement error in it.

Measurement error is the result of random error fluctuations due to the choice of a particular sample of questions or conditions of observation. Esomonu and Okeaba (2021) explained that measurement error occurs when a student's true ability or achievement is either underestimated or overestimated. Measurement error is also seen as the difference between the information in the original state and reporting it in an untrue way. An error of measurement is categorized into random and systematic. Systematic errors also known as bias errors or determinate errors are those errors that consistently affect an individual's score because of some particular characteristics of the person or the things that have nothing to do with the construct being measured which could be caused by hearing impairment (Imasuen & Adeosun, 2023). On the contrary, Imasuen and Adeosun, (2023) posit that random errors are unpredictable variations in measurements that occur due to various uncontrollable factors such as fluctuation in environmental conditions, inherent limitations of the measuring instrument, or human error. These errors can lead to inconsistent and scattered data points when repeated measurements are taken under the same condition.

Sources of random errors as detailed by Imasuen and Ebuwa (2022) include guessing, distractions in the testing situation, administration errors, content sampling, score errors, and fluctuation in the individual examinee's state. Fluctuation in an individual behaviour may be general enough to affect overall test performance, or it may be very brief and specific. If examinees were to repeat the same examination, the random error that affect his/her score on the first occasion probably would not be repeated, although other random errors would undoubtedly occur. Systematic measurement errors cause test scores to be inaccurate and thus reduce their practical utility, while random errors reduce both the consistency and the usefulness of the test score (Iheanyichukwu & Orluwene, 2021). Thus, one of the persistence problem in measurement is how to estimate the degree of error present in a given set of observations. An examinee's score can be expected to vary across different versions of a test because of differences in the way graders evaluate

students' responses and differences in transitory factors such as the student's attentiveness on the day the test was taken. For these reasons, no single test score can be a perfectly dependable indicator of students' ability, hence the need to estimate measurement errors.

Measurement errors can result from the way a test is designed, or from the factors related to individual students or the testing situation, and many other sources like the mood of the examiner, the time of the test (occasion), the test environment, raters, invigilators and the changing order of the questions, which may lead to higher or lower scores (Imasuen & Adeosun, 2023). The need for estimating measurement errors arose because of inconsistencies in measurements, especially those involving multiple sources of error using generalizability theory.

Though several researchers have carried out studies on estimating measurement error in different subject areas, there seems to be a gap in literature. However, to the best knowledge of the researcher, there are limited or no empirical works on estimating measurement error in SSCE economics essay questions conducted by WAEC using generalizability theory where the differentiation facet of students and instrumentation facets of test items and raters or makers are considered. This existing gap necessitated the current study.

Research Questions

The following questions guided the study

1. What is the contribution of the facets: students (s), items (i), and raters (r) and their interaction to measurement error in the 2020 SSCE Economics essay examinations?
2. To what extent do the generalizability coefficients show the degree to which students maintain their rank order across facets of item (i), and raters (r) in the 2020 SSCE Economics essay examinations?

Hypothesis

The following null hypotheses were tested at 0.05 level of significance.

1. There is no significant difference in the contributions of the facets: students (s), items (i), and raters (r), and their interactions to measurement error in the 2020 SSCE Economics essay examinations.

Materials and Methods

The study adopted a random effect design, two-facets fully crossed s x i x r design for a generalizability (G) and decision (D) studies. The researcher used a fully crossed design in the Gstudy so as to estimate all the possible variance components in the measurement situation. The D-study used the G-study's information to design the best measurement procedure in minimizing undesirable sources of measurement error. The population of the study consisted 6,799 Senior Secondary Three (SS3) students from the 137 public Secondary Schools in Okigwe Education Zone for the 2023/2024 academic session. The sample for this study comprised 680 students representing 10% of the total

number of SS3 students in the six Local Government Areas in Okigwe Education Zone. The researcher adopted paper 2 sections A and B of the SSCE in Economics essay paper and its marking guide for this study. The two sections were made compulsory so as to fit into the design of this study which is "students' cross items, cross raters". The section A and B of SSCE Economics comprised all the eight questions in the examination question paper. The instruments are already good questions prepared by WAEC and therefore unnecessary to be subjected to validation and reliability check. Standard error variance was used to answer the research questions. The customary rule for interpretation according to Brennin cited in Huebner and Lucht (2019) was that G- Coefficient equal or superior to 0.80 is evidence that the measure in question was of satisfactory precision. EduG version 6.0-e which was based on the Analysis of Variance (ANOVA) and Generalizability Theory was used to carry out the Generalizability analysis and Decision studies. To test the two hypotheses at 5% significant level, Standard error variance components was used to determine if a significant difference exists in the contributions and effects of the facets to measurement error and score dependability in examination scores. An overlap of the variance components implied that, there is no significant difference but if there was no overlap, then there is significant difference. The justification for this was based on the fact that the ANOVA in Generalizability theory does not compute the F ratio for hypothesis testing but rather it was used to estimate variance components.

Presentation of Results

Research Question 1: What is the contribution of the facets: students (s), items (i), and raters (r) and their interaction to measurement error in the 2020 SSCE Economics essay examinations?

Table 1: Variance Components of the Contributions of the Facets of the Study to Measurement Error in the 2020 SSCE Economics Essay Examination Conducted by WAEC

Sources	Sum of Square	df	Mean Square	Variance Components	Estimates	% of Total Variability	Standard Error
Students (s)	91552.19	679	10172.47	σ^2s	0.20	58.2	8.68
Items (i)	2932.86	49	59.85	σ^2i	0.02	0.0	0.13
Raters (r)	67.80	9	7.53	σ^2r	0.00	0.0	0.01
s x i	27236.58	441	61.76	σ^2si	0.053	15.1	0.42
s x r	656.80	81	8.11	σ^2sr	-0.02	0.0	0.03
r x i	3972.97	441	9.01	σ^2ri	-0.03	0.0	0.06
s x i x r	36918.13	3969	9.30	σ^2sir	0.09	26.7	0.21
Total	**163337.33**	**5669**				**100%**	

Table 1 shows that the highest contribution to measurement error in the 2020 SSCE Economics essay examinations scores came from the students residual σ^2s 0.20; accounting for (58.2%) of the total variability. This was followed by the

interaction of student, items and raters factors σ^2sir 0.09; which accounted for 26.7% of the total variability. The third highest contributor to measurement error was the interaction of students and items factors σ^2si 0.053; which accounted for 15.1% of the total variability. The facets items and the interaction of the students and raters, and raters and items did not contribute to measurement error in the study. This was an indication that measurement error exists in the examination scores with the object of measurement (students) contributing the highest of the total variability in the 2020 SSCE Economics essay examination scores.

Research Question 2: To what extent do the generalizability coefficients show the degree to which students maintain their rank order across facets of item (i), and raters (r) in the 2020 SSCE Economics essay examinations?

Table 2: Estimated Generalizability Coefficients ($E\rho^2$) for a Fully Crossed SxRxI D-Study Design with Different Levels of Items and Raters

		SSCE 2020			SSCE 2020
Raters		$E\rho^2$	**Item**		$E\rho^2$
Initial	5	.72	Initial	8	.72
	6	.74		9	.81
	7	.76		10	.82
	8	.79		11	.84
	9	.81		12	.87
	10	.83		13	.90

Table 2, shows the impact of increasing the levels of raters and items on students (objects of measurement) in terms of their relative standing. They showed a steady but gradual increase. When the number of raters was increased from 5 to 6, generalizability coefficient ($E\rho^2$) was increased to 0.74 for WAEC 2020. A further increase in the level to 7 raters brought about a 0.76 increase in generalizability coefficient of SSCE 2020. At the level of 10, ($E\rho^2$) increased to 0.83 generalizability coefficient of SSCE 2020. At this point, students are ranked ordered relatively. This showed that an increase in the number of raters rank ordered students according to their relative ability in examination scores.

Hypothesis 1: There is no significant difference in the contributions of the facets: students (s), items (i), and raters (r), and their interactions to measurement error in the 2020 and 2021 SSCE Economics essay examinations conducted by WAEC.

Table 7: 95% Confidence Interval on G-Study Variance Components

	SSCE 2020	
Components	**Lower limit**	**Upper limit**
Students (S)	.959	1.035
Raters (R)	-.884	1.112
Item (I)	-.947	1.053
s x i	-.918	1.076
s x r	-.999	1.001
r x i	-.948	1.052
s x i x r	-.865	1.135

* Significant at 0.05 level of significance

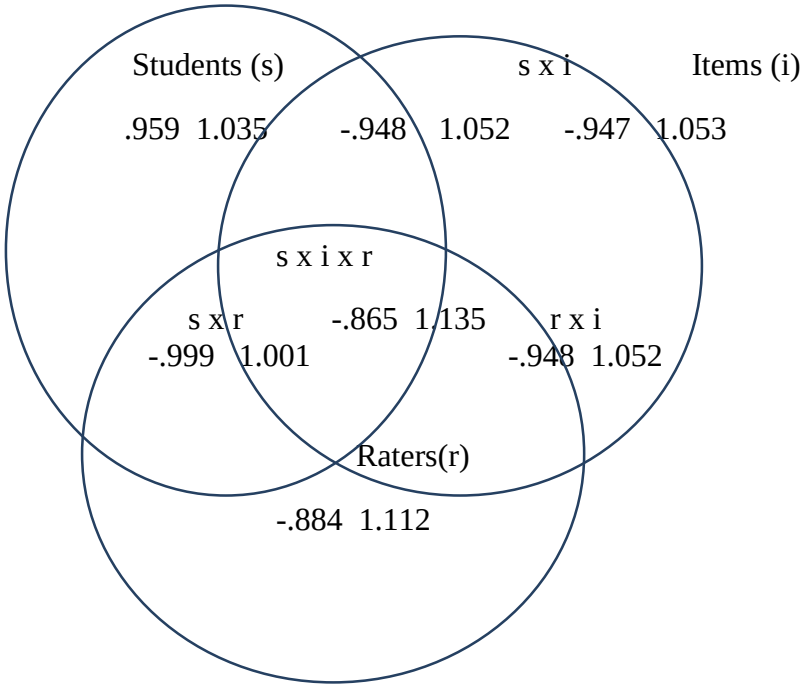


Fig 1: Venn Diagram Showing the Overlap Variance Components of the Facets of Measurement Error in SSCE 2020 SSCE Economics Essay Examinations

Table 3, Figure 1 show that the variance of students (σ^2s) did not overlap indicating that their variance components were significantly different in their contributions to measurement error. But the items (σ^2i), raters (σ^2r), the interaction of students and items (σ^2si), students and raters (σ^2sr) variance components overlapped, indicating that their variance components were not significantly different in their contributions to

measurement error in the 2020 SSCE Economics essay examinations conducted. Therefore, the null hypothesis which states that there is no significant difference in the contributions of the facets: students (s), items (i), and raters (r), and their interactions to measurement error in the 2020 SSCE Economics essay examinations was retained.

Discussion of Findings

The findings from research question one shows that the value of the relative generalizability coefficient was 0.91, and the absolute generalizability coefficient was 0.90, suggesting that the measurement design and procedure with the current number of students, raters and items were satisfactory. This is in agreement with the findings of Cardinet, *et al* (2011) that in educational assessment context, either the students, the interaction between students and questions most times is the highest or most important contributor to measurement error. The G-coefficient and dependability coefficient for the items were 0.90 and 0.89 respectively, an indication that the items were reliable and valid. This is in consonant with the rule of thumb that a generalizability coefficient above 0.80 showed satisfactory reliability for any given measurement (Esomonu & Okeaba, 2021). This high reliability of the scores was in tandem with the findings of Akindahunsi and Afolabi (2021) whose study showed that generalizability coefficient of 0.90 and phi coefficient of 0.87, was an indication that a high variability of scored was obtained.

The findings of research question one also revealed that the highest contributors to measurement error was the students facet that accounted for 58.2% in 2020 SSCE economics examination conducted by WAEC. This agreed with Arterberry, *et al*. (2021) whose study revealed that the facet person had the highest variance to measurement error. It was also in agreement with Ene, *et al*. (2021), Marty, *et al*. (2010), and Lee (2010) whose studies revealed that the examinees (students) contributed the highest variance to measurement error. The findings from research question three showed an increase in the number of raters to 10 and items 13, rank-ordered students according to their relative standing or ability in the examinations. A change in the number of raters from 5 to 10 and items from 8 to 13 for 2020 SSCE economics examination increased the generalizability coefficient to 0.86. Thus, increasing the number of raters to 10 and items to 13 would improve the reliability of the data, but will require a large increase in resources to collect the data (Kline, *et al*, 2020). This is in agreement with Imasuen and Adeosun (2023), Esomonu and Okeaba (2021), Fulcher (2003), and Lee, *et al* (2001).

This result was also in conformity with the findings of Web, *et al* (2006) who opined that more observers gave generalizability coefficient that ranked-ordered person in the generalizability study of job performance measurement of Navy Machinist Mates. In the same vein, the result agreed with Brennan (2001) whose study revealed that with multiple raters, it was possible to differentiate between persons. Therefore, to achieve a high generalizability

coefficient, the number of raters needs to be increased to rank order students relatively in examinations. The study also corroborated with the study of Arterberry, *et al* (2021) which revealed that the reliability of scores indicated a fairly large improvement when increasing the number of items. It was also in consonant with the study of Lombardi, *et al* (2021) which revealed that at least six raters did reach an acceptable level of dependability in estimating Mathematics and English students' level of cognitive domain.

The findings of the study in hypothesis one showed that the variance components of students did not overlap, indicating that their variance components were significantly different in their contribution to measurement error. But the items, raters, the interaction of students and items, students and raters, raters and items, and the residual, variance components overlapped, indicating that their variance components were not significantly different in their contributions to measurement error in the 2020 SSCE economics examinations scores. This was in line with the findings of Esomonu and Okeaba (2021) whose study revealed that the variance components were not significantly different in their contribution to ISIUAC scores. This result corroborated the study of Egbulefu (2013) who stated that instrumentation facet did not differ significantly in their contribution to measurement error in the examination.

Conclusion

From the analyses of data collected and results presented, it could be concluded that the students and interaction between the facets contributed to measurement error in the SSCE in economics for 2020. The students' facets were significantly different in their contributions to measurement error ($p<0.05$) while the other facets and their interactions were not significant.

Recommendations

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

1. Generalizability analysis should be carried out by test developers, psychometricians, policymakers in assessment in the estimation of reliability to estimate multiple sources of error and to reduce or eliminate measurement error and hence maximize reliability of students' scores.
2. More works should be done by the teacher in the classroom with the students in preparing them for external examinations since the student facet and its interaction with items contributed substantially to measurement error.

REFERENCES

Adeosun, P. K. (2020). Assessment of model-fit of 2019 BECE Mathematics multiple-choice items in secondary schools in Benin metropolis. *Nigerian Journal of Applied Psychology*, 22:214 - 227

Akindahunsi, O. F., and Afolabi, E. R. I. (2021). Using generalizability theory to investigate the reliability of scores assigned to students in English language examination in Nigeria. *Journal of Measurement and Evaluation in Education*

and Psychology, 12(2), 147-162. doi: 10.2103 1/epod.820989.

Arterberry, B. J., Martens, M.P., Cadigan, J. M. & Rohrer, D. (2021) Application of Generalizability Theory to the Big Five Inventory. *Personality and individual differences* 69(1), 98-112.

Brennan, R.L (2019). *Generalizability Theory*: Statistics for Social Science and Public Policy. Springer-Verlag Berlin Heidlberg. New York.

Egbulefu, C.A. (2013). Estimating measurement error and score dependability in examinations using generalizability theory. *(Unpublished doctoral dissertation).* University of Nigeria, Nsukka.

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 Generalizability Theory. *Rivers State University Journal of Education (RSUJOE)*, 24 (1):35-46

Esomonu, N.P.M & Ikeanumba, C.B. (2021). Effect of formative assessment on academic achievement of postgraduate students in advanced educational statistics in public universities in southeast, Nigeria. *International Journal of Advanced Education and Research, 6(4), 05-11.*

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

Hueber, A. & Lucht, M. (2019). *Generalizability Theory in R*. Practical Assessment, Research 24(1) and Evaluation, 24(5) 1-12. https://doi.org/10.7275/5065.gc10 RSUJOE 2021,

Imasuen, K. & Ebuwa, S.O. (2022). Assessing score dependability of West Africa Examination Council (WAEC) 2019 mathematics objective test using generalizability theory. *British Journal of Cotemporary Education*, 2(1), 64 -73.

Lee, Y., Kantor, R., & Mollaun, P. (2002). *Score dependability of the writing sections of new TOEFL*. Paper presented at the Annual Meeting of National Council on Measurement in Education, New Orleans, LA

Lombardi, A. Seburn, M., Conley, D., & Snow, E. (2021). *A Generalizability Investigation of Cognitive Demand and Rigor Ratings of Items and Standards*. Educational Policy Improvement Center Eugene, Presented at the annual conference of the American Educational Research Association Denver.

Omorogiuwa, O.K. (2019). *An Introduction to Educational Measurement and Evaluation*. Benin City: Mase-Perfect.