**International Journal of Advanced Education and Research**
**www.multidisciplinaryjournals.net**
**ISSN: 2455-5746**
**Received: 12-04-2024, Accepted: 03-05-2024, Published: 28-05-2024**
**Volume 9, Issue 2, 2024, Page No. 12-15**

# Estimating the effects of the instrumentation facets and their interractions on score dependability in examinations using generalizability theory

**Chukwuemeka B Ikeanumba, Nkechi P M Esomonu**
Department of Educational Foundations, Faculty of Education, Nnamdi Azikiwe University, Awka, Nigeria

**Abstract**
This study estimated the effects of instrumentation facets and their interactions on score dependability in examinations, using the Generalizability Theory. With students' low performance in economics examinations. It is needful to estimate the effects of the instrumentation facets and their interaction on score dependability using generalizability theory. Two research questions and one hypothesis were posed to guide 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 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. Hypothesis was 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 dependability study showed that SSCE in economics for 2020 will be optimized using 9 raters to 13 items. 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, dependability, and students

## Introduction

Examinations are frequently utilized to determine students' placement within a classroom setting or to assess their level of proficiency within a particular subject area, as well as for certification purposes. The assessments' scores are utilized to assess students. However, a hurdle arises due to the presence of errors in these scores, thereby making the scores achieved in exams an inaccurate reflection of students' true abilities. As stated by Esomonu and Okeaba (2021) [4], test scores do not provide a conclusive measure of a student's knowledge or skills. It is anticipated that an examinee's score may differ across various versions of a test. Such score variance frequently arises from variations in how markers assess students' responses and differences in transient factors like the student's attentiveness on the day the test was taken, student's health on the day test was taken, student's negative attitude towards the subject, among others. For this reason, no individual test score can serve as an entirely reliable indicator of a student's performance.

As viewed by Johnson, Dulany & Banks in Esomonu and Okeaba (2021) [4], random errors in measurement may arise from various sources, including test design, individual student characteristics, testing conditions, and other sources such as examiner mood, timing of the test (occasion), testing environment, invigilators, and alterations in the sequence of questions. These sources can potentially influence scores to be either higher or lower. Certain test items (questions) might exhibit bias either in favor of or against specific student groups. The necessity to estimate measurement error arises due to inconsistencies in measurements and the frequency at which students fail in examinations.

According to Chief Examiner's Report (2023), the result of the 2016-2019 May/June WAEC revealed that in 2016, out of 1,540,902 candidates representing 97.29% that wrote Economics only 864,273 representing 56.09% had credit while 641,789 candidates representing 41.65% failed. In 2017, 1,025,703 candidates representing 66.94% had credit while 470, 890 candidates representing 30.74% failed to obtain credit. Furthermore, in 2018, there was a decrease in performance as out of 1,363,994 candidates representing 98.05% that sat for the examination, only 698,669 representing 51.22% had credit while 639,086 candidates representing 46.85% failed to have credit in the subject. The same decrease in performance occurred in 2019 as 511,007 candidates representing 43.47% had credit while 639,153 candidates representing 54.37% failed the subject. Furthermore, the WAEC result released in 2020 reflected that a total of 1,312,414 candidates sat for WAEC economics, the result shows that 43.82% had credit in the subject while 56.18% failed to do so. The same failure occurred in WAEC 2021 where 24,722 candidate representing 41.19% passed with credit while 58.81% failed. The question at hand is whether these scores accurately represent the performance of students in the examination? The observed low performance of students prompts the need to estimate multiple sources of error to determine the contributions of various examination facets to overall error. By doing so, we can identify methods to minimize or eliminate these errors, consequently increasing the reliability of examination scores.

Scores obtained in examinations can be influenced by factors beyond the cognitive ability of a student. These factors, including test questions, invigilators, and raters, are

probable influencers on the reliability of observed scores in exams, consequently affecting their interpretation and subsequent decision-making (Esomonu & Okeaba, 2021) [4]. The ramifications of these influences raise concerns regarding the accuracy, precision, and fairness of students' examination scores.

Estimating measurement error and score reliability in examinations involves a multiple facet approach, therefore the Classical Test Theory which has been widely used before now is not suitable to be used in assessing the effects of multiple sources of error because it focuses on one source of measurement error per time. On this premise, this study seeks to estimate the effect of instrumentation facets on measurement error and score dependability using Generalizability theory. More so, in the generalizability theory, Instrumentation facets are, "instruments" that you use to collect the quantitative information, where "instruments" embraces both measurement tools, principally the test questions, and measurement procedures, such as conditions of observation, rules for interpreting the answers, raters, questions, procedures, conditions, and rules for scoring, among others. Since instrumentation facets and their interactions contribute to measurement error, it is needful to design a study to estimate its effect on score dependability using generalizability theory.

**Research Questions**

The following questions guided the study

1. What is the contribution of the instrumentation facets: items (i), and raters (r) and their interaction to score dependability in the SSCE Economics essay examinations?
2. To what extent do the dependability coefficients show the degree to which students maintain their rank order across facets of item (i) and raters (r) in the 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 instrumentation facets: items (i), raters (r), and their interactions to score dependability in the 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 and maximizing reliability. This is represented in the Venn diagram in figure 1. The circles represent the facets; students, raters (r) and the test questions items (I). Circle overlap areas represent facet interactions, and the seven distinct areas correspond to the seven effects, $\sigma^2 s$, $\sigma^2 i$, $\sigma^2 r$, $\sigma^2 si$, $\sigma^2 sr$, $\sigma^2 ri$, *and* $\sigma^2 sir$.

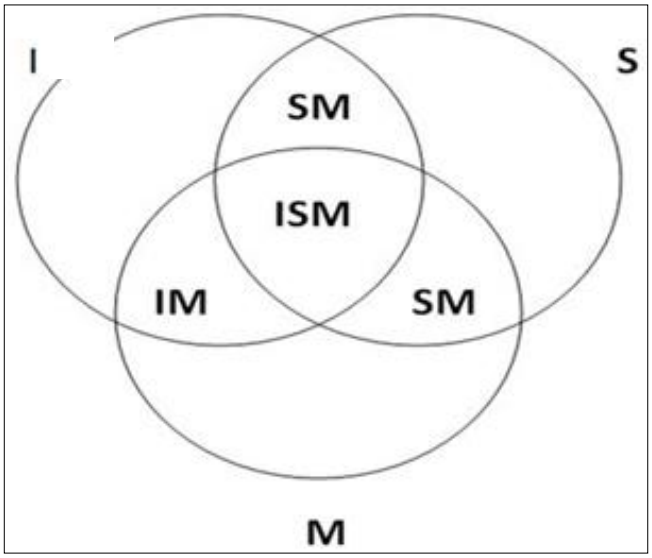


**Fig 1**

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.

## Results

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

**Table 1:** Effects of Items, Raters, and their Interactions to Score Dependability in the SSCE Economics Essay Examination

| Source | Variance Component Estimates | Relative Error Variance | % Relative Error Variance | Absolute Error Variance | % of Absolute Variance |
|---|---|---|---|---|---|
| Students (S) | 20.224 | …….. | …….. | …….. | …….. |
| Raters | | …….. | …….. | (0.000) | 0.0 |
| Items | | …….. | …….. | (0.000) | 0.0 |
| s x i | | 0.105 | 84.9 | 0.103 | 84.9 |
| s x r | | (0.000) | 0.0 | (0.000) | 0.0 |
| r x i | | …….. | …….. | (0.000) | 0.0 |
| s x i x r | | 0.019 | 15.1 | 0.017 | 15.1 |
| Total | | 0.124 | 100% | 0.12 | 100% |

Error Variances $\sigma^2\delta = 0.124$; $\sigma^2\Delta = 0.124$ Coefficients: $E\rho^2 - 0.90$, Ø – 0.89.

Table 1 show that the absolute error variance for items, raters, the interaction of items and students, raters and items were set at zero. Conversely, the absolute error variance estimate for the interaction of students and items was 0.103 accounting for 84.9%, while the absolute error variance estimate for the interaction of the students, items and raters was 0.017, accounting for 15.1%. The dependability index (Ø) of 0.89, showed that the 5 raters were similar and consistent in their rating, thereby, yielding a high dependability index. Also, the variance component estimate of 20.224 is an indication that the scores were valid.

**Research Question 2:** To what extent do the dependability 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 examination?

**Table 2:** Estimated Dependability Coefficients (Φ) for a Fully Crossed S x I x R D-Study Design with Different Number of Raters and Items

| | SSCE 2020 | | SSCE 2020 |
|---|---|---|---|
| **Raters** | **Φ** | **Item** | **Φ** |
| Initial 5 | .73 | Initial 8 | .73 |
| 6 | .75 | 9 | .78 |
| 7 | .85 | 10 | .81 |
| 8 | .89 | 11 | .82 |
| 9 | .92 | 12 | .84 |
| 10 | .94 | 13 | .95 |

Table 2 show that with 5 raters the dependability index (Φ) was 0.73 for SSCE 2020 in economics. When the number of raters was increased to 6, the dependability index (Φ) was 0.75 for WAEC 2020 in economics. This is not high enough to separate students in terms of performance. However, the potential effect of the change in measurement reliability was felt by increasing further the number of raters to 10. An increase in the number of raters to 10 produced dependability index (Φ) increase of .94 for SSCE 2020 in economics. This is high enough to classify students (the object of measurements) in terms of performance, irrespective of the performance of others.

**Hypothesis 1:** There is no significant difference in the contributions of the instrumentation facets: items (i), raters (r), and their interactions to score dependability in the 2020 SSCE Economics essay examinations.

**Table 3:** 95% Confidence Interval on D-Study Variance Components

| | SSCE 2020 | |
|---|---|---|
| **Raters** | **Lower limit** | **Upper limit** |
| students (S) | -1.000 | 1.000 |
| Raters (R) | -.993 | 1.007 |
| Item (I) | -.998 | 1.002 |
| s x r | -.972 | 1.028 |
| s x i | -1.000 | 1.000 |
| r x i | -.998 | 1.002 |
| s x i x r | -.994 | 1.006 |

* Significant at 0.05 level of significance

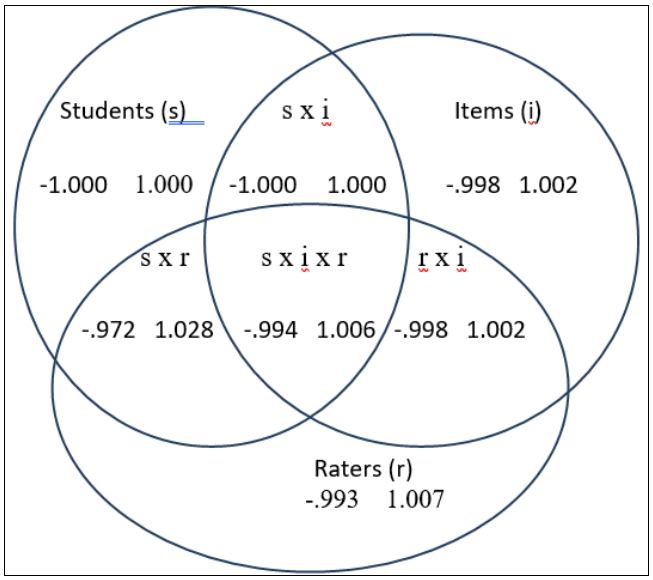


**Fig 2:** Venn Diagram Showing the Overlap Variance Components of the Facets on Scores Dependability in 2020 SSCE Economics Essay Examinations

Table 3, Figure 2 that the variance components for the instrumentation facet (items, raters and their interactions) overlapped. This was an indication that the instrumentation facets did not significantly differ in their contribution to score dependability in the 2020 SSCE economics examinations. Therefore, the null hypothesis which states that there is no significant difference in the effects of the instrumentation facets: items (i), raters (r), and their interactions to score dependability in the 2020 SSCE Economics essay examinations was retained at (p>0.05).

**Discussion of Findings**

The findings from research question one revealed that the highest effects on score dependability in 2020 SSCE economics examination came from the interactions of students which accounted for 84.9% on score dependability. This was followed by the interaction of students, items and raters which accounted for 15.1% for 2020 on score dependability in SSCE economics examination. Items and interaction of student and raters for 2020; and items and interaction of students and items did not have any effect on score dependability in the examinations scores. This is because the raters who scored the students maximized their observed scores in the examination. Also, it was observed that more of the absolute error variability in the examination came from raters, thus, changing the level (numbers) of the raters will have a large effect on the score dependability than changing the number of items. Therefore, there will be the need to bring in more raters to bring about dependable scores in the examination. These findings in the study were consistent with the findings of Imasuen and Ebuwa (2022) [6], Fulcher (2003) and Lee, *et al*. (2001).

For research question two, the level of raters at 5 and items at 8 was not quite satisfactory to produce an absolute scale of measurement. There should be at least 10 raters and 13 items to attain a dependability index of 0.95 for 2020 SSCE economics examinations conducted by WAEC respectively will help to successfully separate students in terms of their performance irrespective of the performance of other students. The results were consistent with the study of Brennan (2001) [2] who found that more raters were needed for a high dependability index. The findings of the study was also supported by the findings of Esomonu and Okeaba (2021) [4] that an increase in the number of markers yielded a higher dependability index than when the raters were small in a study on the dependability of score. The finding is also in tandem with the findings of Arterberry, *et al* (2021) [1] whose study showed that the reliability of scores indicated a fairly large improvement when increasing the number of items.

The findings in hypothesis one showed that the variance components for the instrumentation variance (items, raters and their interactions) overlapped. Therefore, the null hypothesis was retained. This was an indication that the instrumentation facets did not differ significantly in their interaction to score dependability in the examinations scores. This was in consonance with the findings of Egbulefu (2013) [3] whose study revealed that the instrumentation facets though contributed to score dependability but their contribution was not significantly different. This was in agreement with the findings of Esomonu and Okeaba (2021) [4] in which markers were not significantly different in their contribution to score dependability. However, the item was found to be statistically significant.

**Conclusion**

From the analyses of data collected and results presented, it was revealed that the highest effects on score dependability in examination came from the interaction between students and items for 2020 and apart from the items facet, other sources (facets) affect the scores students obtained in the examinations. 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. An increase in the level of items and increase in the level of raters to yielded a very high dependability coefficient index.

**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. Psychometric Units for examination bodies such as WAEC and NECO should ensure that the items are subjected to the process of test development. Item writers should write items that will help to distinguish between students of different achievement levels. This will reduce error in measurement and ensure score dependability.
3. There should be enough raters when scoring students' performance in examinations. This will help teachers in increasing the objectivity of scoring, minimizing error and maximizing reliability of examination scores.

**References**


1. Arterberry BJ, Martens MP, Cadigan JM, Rohrer D. Application of Generalizability Theory to the Big Five Inventory. Personality and individual differences,2021:69(1):98-112.
2. Brennan RL. *Generalizability Theory*: Statistics for Social Science and Public Policy. Springer-Verlag Berlin Heidlberg. New York, 2001.
3. Egbulefu CA. Estimating measurement error and score dependability in examinations using generalizability theory. *(Unpublished doctoral dissertation).* University of Nigeria, Nsukka, 2013.
4. Esomonu NP, Okeaba JU. 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),2021:24(1):35-46.
5. Hueber A, Lucht M. *Generalizability Theory in R*. Practical Assessment, Research and Evaluation,2019:24(5):1-12.
6. Imasuen K, Ebuwa SO. Assessing score dependability of West Africa Examination Council (WAEC) 2019 mathematics objective test using generalizability theory. British Journal of Cotemporary Education,2022:2(1):64 -73.
7. Lee Y, Kantor R, Mollaun P. 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, 2002.
8. West Africa Examination Council Chief Examiners' Report. General Comment, weakness/remedies and strength. *Retrieved from,* 2023. *https://waeconline.org.ng/elearning/Biology/Bio227mc.html*