I. INTRODUCTION

Measurement error is a situation in which a student’s true ability or achievement is either underestimated or overestimated [1]. According to [2], measurement error is defined as the difference between the distorted information and undistorted information about a measured product expressed in its physical quantity. [3] noted that under an ideal condition, measurement has no error, but that ideal conditions are difficult to put in place.

Regarding measurement error, it is important to emphasize that scores of tests are either reliable or unreliable [4], [5]. An examinee’s score can be expected to vary across different versions of a test because of differences in the way markers assess student’s responses and differences in transitory factors such as the student’s attentiveness on the day the test was taken, student’s health on the day test was taken, etc. For these reasons and more, no single test score can be a perfectly dependable indicator of a student’s performance hence test scores are not always a definitive measure of student’s knowledge or skills.

Although all measurements are fallible to some extent, evaluators seek ways to increase the precision of measurement. To do so, they frequently average measurements over some subset of predefined conditions of measurement. This average measurement serves as an estimate of the “ideal” measurement that would be obtained (hypothetically) by averaging over all predefined conditions.
of measurement [6]. A substantive question then becomes, “how many times would one measure to determine the average? Another way that evaluators sometimes increase the precision of measurement is to fix one or more measurement conditions [6]. For example, a specific ruler might be used to obtain all measurements of an object’s length. However, a specific ruler’s choice for all measurements involves a restriction on the set of measurement conditions to which generalization is intended. In other words, fixing a condition of measurement reduces error and increases the precision of measurements, but it does so at the expense of narrowing interpretations of measurements. It is, therefore, evident from this perspective of measurement that “error” does not mean mistake, and what constitutes error in a measurement procedure is, in fact, a matter of definition. It is one thing to say that error is an inherent aspect of a measurement procedure; it is quite another thing to quantify error and specify which conditions of measurement contribute to it. Doing so necessitates specifying what would constitute an “ideal” measurement (i.e., over what conditions of measurement is generalization intended) and the conditions under which observed scores are obtained [7]. Therefore, these conditions need to be stated so the contributors to error in measurement can be identified for possible reduction or elimination.

More so, for subjects like mathematics whose relevance is so much, it is necessary to estimate measurement errors. Despite the significance of the subject, [8] noted that over the years, the performance of students in senior secondary school mathematics in Nigeria have consistently been poor and unimpressive noting that available data from the two public examination bodies, i.e., (the West African Examination Council [WAEC] and National Examination Council [NECO]) Secondary School Certificate Examination (SSCE) indicate that students’ achievement in mathematics is low. For example, about 71% of the candidates who sat for the October/November 2014 West African SSCE private failed to obtain five credits with English Language and Mathematics. A total of 241,161 candidates sat for the examination, but only 72,522 candidates representing 29.37% were successful in getting credits in five subjects that include English Language and Mathematics [9]. The West African Examinations council results in 2018 reflected that a total of 1.57m candidates sat for WAEC as public students, the results shows that 48.15% had 5 credits and above including English and Mathematics while 51.85% failed to do so. In the same year a total of 109,798 candidates sat for WAEC as private students but only 33.81% of them had 5 credits and above including English and Mathematics while 66.19% did not [10].

This low performance of students in examinations calls for the estimation of multiple sources of error, to determine the contributions of the different facets in examination to error and then see how these errors can be minimized or eliminated and hence increase reliability of examination scores. Generalizability theory, or G theory, is essentially an approach to the estimation of measurement precision in situations where measurements are subject to multiple sources of error. It is an approach that not only provides a means of estimating the dependability of measurements already made, but also enables information about error contributions to be used to improve measurement procedures in future applications. The originality of the G theory approach lies in the fact that it introduced a radical change in perspective in measurement theory and practice. In essence, the classical correlational paradigm gave way to a new conceptual framework, deriving from the variance analysis (ANOVA), whose fundamental aim is to partition the total variance in a data set into several potentially explanatory sources. Despite this profound change in perspective, [11] noted that G theory does not in any way contradict the results and contributions of classical test theory. But rather, what classical test can estimate one per time (stability, equivalence, internal consistency, validity, inter-rater agreement, etc.), G theory brings together as one. That is, the various reliability coefficient that classical test estimates can be put together and estimated as generalizability and dependability coefficients in G theory. The impact of the change in perspective is more than a straightforward theoretical reformulation. The fact that several identifiable sources of measurement error (markers, items, gender, etc.) can simultaneously be incorporated into the measurement model and separately quantified, means that alternative sampling plans can be explored with a view to controlling the effects of these variables in future applications [11].

Furthermore, the joint standards recommend that: when an instrument is used to make group judgments, reliability data must bear directly on the groups’ interpretations. Standard errors appropriate to individual scores are not appropriate measures of the precision of group averages. A more appropriate statistic is the standard error of the observed score means. Generalizability theory can provide more refined indices when the sources of measurement are numerous and complex [12].

Despite these recommendations, researchers still rely on classical test theory for calculating instrument reliability. [13] noted that the essential difference between measurement error as conceived in the a coefficient and measurement error as conceived in a more complex G coefficient is that in the former case measurement error is attributable to one single source of variance, the student-by-item interaction that is the inconsistent performances of individual students over the items in the test, whereas in the latter case multiple sources of error variance are acknowledged and identified. He further noted that a G coefficient of relative measurement indicates how well a measurement procedure has differentiated among objects of study, in effect how well the procedure has ranked objects on a measuring scale, where the objects concerned might be students, patients, teaching methods, training programs, or whatever. This is also what the a coefficient does, but in a narrower sense. There is therefore need to develop a study to estimate dependability of test scores using generalizability theory as this would further the campaign for the use of generalizability theory over the use of classical test theory in the estimation of reliability of test scores and when applied to test development, improve students’ performance in mathematics. This present study seeks to employ generalizability theory to estimate measurement error and score dependability of test scores in NECO mathematics using students, test questions and markers as facets.
II. MATERIALS AND METHODS

A. Population and Sample

The population for the study was all the 5,085 SS3 students of the 34 Government owned senior secondary schools in Yenagoa LGA of Bayelsa State. (Bayelsa State Post Primary School Board 2019). The senior secondary 3 students were considered best for the study, on the grounds that they were the only class ready for the NECO External Examinations and were expected to have covered the required syllabus for Mathematics. The sample for the study was 1,525 students. This is approximately thirty percent (30%) of the population of the study. 10 public senior secondary schools were selected through simple random sampling of Balloting without replacement. All students in the selected schools will form the sample for the study.

B. Instruments

The researcher adopted section A of the 2019 NECO Mathematics main paper and 2019 NECO Mathematics Marking Scheme for this work. The section A of NECO 2019 mathematics, comprised of five compulsory question with eight marks each drawn from logarithm, indices, algebraic expressions, circle theory and descriptive statistics. Items 1a, 2 (a & b), 3b, and 5b measure application, Items 1b and 4b measure analysis while items 4a and 5a measure measure.

This section only is used because the section “B” has optional questions and will not fit into the design of this study which is “students’ cross question, cross makers”. The instruments are already good questions prepared by NECO and therefore unnecessary to be subjected to validation and reliability check.

C. Data collection:

The students were given the NECO Mathematics main paper during their MOCK Examination so as to instill seriousness. The students were graded by 3 markers using the NECO Marking Scheme. To eliminate markers’ bias, 3 copies of each student’s scripts was made and distributed to the markers.

D. Data Analysis

EduG version 6.0-e which is based on the Analysis of Variance (ANOVA) and Generalizability Theory were used to carry out the Generalizability analysis. It was used to answer the 3 research questions. To test the hypothesis at 5% significant level, standard error variance components were computed 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 will imply that, there is no significant difference but if there is no overlap, then there is significant difference. The justification for this is based on the fact that the ANOVA in Generalizability theory does not compute the F ratio for hypothesis testing but rather it is used to estimate variance components.

III. RESULTS

Tables I and II showed that the highest contribution to measurement error in examination scores comes from the students’ facet (σ2S) 4.13352, accounting for (53.1%) of the total variability. This was followed by the residual σ2SM (1.6369; accounting for 21.0% of the total variability. The interaction of student x items (σ2SI) accounted for 17.9% while the interaction of students x markers (σ2SM) accounted for 1.4%.

Table III showed the impact of increasing the level of markers on students while the level of items is stable. They showed a steady but gradual increase. When the level of markers increased from 1 to 2, and the level of items is 5 for both markers, generalizability coefficient (EP2) was increased by 0.04. A further increase in the level of markers to 4, while we still have our level of items at 5, brought about a drastic 0.06 increase (0.85 to 0.91). However, when the level of makers is reduced to 2 and the level of items increased to 10, a very high generalizability coefficient of 0.94 was obtained.

| Source | SS        | MS        | Variance component | Estimate | DF | % Of total variability |
|--------|-----------|-----------|--------------------|----------|----|------------------------|
| Students (S) | 84128.112 | 68.341    | σ2S                | 4.13352  | 1231 | 53.1%                  |
| Items (I)  | 7292.041 | 1823.010  | σ2I                | 0.48160  | 4   | 2.6%                   |
| Markers (M)| 55.756 | 27.87814  | σ2M                | -0.00201 | 2   | 0.0%                   |
| S x I      | 28593.825| 5.80703   | σ2SI               | 1.39004  | 4924 | 17.9%                  |
| S x M      | 5338.644 | 2.16842   | σ2SM               | 0.10630  | 2462| 1.4%                   |
| M x I      | 318.752  | 39.844    | σ2IM               | 0.03133  | 8   | 0.4%                   |
| S x M x I  | 16120.181| 1.6369    | σ2SM               | 1.6369   | 9848 | 21.0%                  |

TOTAL: 141847.312

18479 100%

| Source | Vc Estimate | Relative Error Variance | % Absolute Error Variance | % Absolute Error Variance |
|--------|-------------|-------------------------|---------------------------|---------------------------|
| Students (S) | 4.13352 | 0.09632 | 18.5 | |
| Items (I) | | | | |
| Markers (M) | (0.00000) | 0.0 | | |
| S x I | 0.27801 | 65.8 | 53.4 | |
| S x M | 0.035443 | 8.4 | 6.8 | |
| I x M | 0.00209 | 0.4 | | |
| S x I x M | 0.10913 | 25.8 | 20.9 | |

TOTAL: 0.42257 100% 0.52098 100%

Coefficients: EP2 – 0.91, Φ = 0.89.
Table IV showed that with 1 marker the dependability index (Φ) was 0.84. When the level of markers was increased to 2 and 5 level of item, the dependability index (Φ) was 0.87, an increase of 0.03. However, setting in the level of markers to 2 with 10 level of items, produced an increase of 0.09 (0.84 to 0.93) in dependability index. This is high enough to classify students in terms of performance, irrespective of the performance of others. This showed that the performance of an individual student does not affect the performance of another student in the given test.

**Hypothesis one:** There is no significant difference in the contributions of facets: students (s), questions (q), markers (m), and their interactions to measurement error in examination scores

Table V, Fig. 1 showed that students (σ²S) variance components did not overlap, indicating that their variance components were significantly different in their contributions to measurement of error in examination scores. The other facets variance components overlapped, indicating that the variance components were not significantly different in their contributions to measurement error in examination scores. Ho1 for facets students will not be accepted, while Ho1 for other facets will be accepted

**IV. DISCUSSION**

The findings from this study showed that the highest contribution to measurement error in examination scores was the student facets (σ²S) 4.13352, accounting for (53.1%) of the total variability. This finding of this result means that factors relating to students greatly affected the performance of the student in the exam. The findings were in line with that of [1] who reported that error can result from the way the test is designed, factors related to the individual students, the testing situation, and many other sources. According to [1], factors relating to students includes fatigue, stress from home, hunger etc., the second largest source of variation to measurement error in examination scores was the residual (σ²SMI), 1.6369; accounting for 21.0% of the total variability. This showed that the residual contributed substantially to the measurement error in examination scores which is line with the findings of [14] who also reported the residual effect as a major source of measurement error, this finding also found support in [15], whose study also found the residual as one of the major contributors to measurement error. This study's result was also supported by the findings of [6] who reported that the residual also made the highest contribution to measurement error. The study of [16] which equally reported that the largest variance was accounted for by the residual is similar to the findings of this study. Apart from the observed facets in these studies, the residual represents other facets that were not included in the study; and these facets (residuals) contribute substantially to error variance. The fact that several identifiable sources of measurement error (markers, items, gender, schools, teacher qualification etc.) can simultaneously contribute to measurement error. The American Educational Research Association (AERA) recommend that where feasible, the error variances arising from each source should be estimated [12].

Furthermore, the findings of this study showed that students variance components did not overlap, indicating that their variance components were significantly different (p < 0.05) in their contributions to measurement error in examination scores. However, for other facets: items, markers, residual and the interactions of students’ x markers
items x markers, their variance components overlapped, hence they were not significantly different. The result of the hypothesis is in agreement with that of [17], [18] and [6] who found the facet students’ significantly different in their studies while other facets apart from the residual were not significantly different in their contribution to measurement error. From the findings, the interactions of students’ x items which ought to be the most important contributions to measurement error in an educational context, contributed very little to measurement error in examination scores. This showed that the magnitude of an effect is not inferred from statistical significance in generalizability theory.

Results further revealed that, a change in the level of markers and items, led to increase or decrease of the generalizability coefficient. With the level of markers at 4 and the level of items at 10, a generalizability of 0.93 was obtained which is high enough to rank students according to their relative standing or ability in examinations. The changing level of markers and items produced a larger effect on the estimated generalizability coefficient (EP2). A change level of items and markers from 5 to 10 and from 3 to 2 respectively increased the generalizability coefficient (EP2) by 0.08 (0.85 to 0.93), when the level of markers was change to 5 while we have our level of items at 15 the (EP2) increased drastically to 0.96. This result was in conformity with the findings of the study by [19] whose study showed that with multiple raters, it was possible to differentiate between persons. The result was also in agreement with the findings of the study by [20] in which more observers yielded a generalizability coefficient that rank ordered persons in a generalizability study of job performance measurements of Navy Machinist mates. Similarly, this result conformed to [6] findings, which revealed that an increase in level of invigilators produced a large effect on the estimated generalizability coefficient.

We found that with a dependability index of 0.93, students who passed were comfortably separated from those who failed. The students who attained the pass score, or above it was separated from the students who did not perform well. The level of markers at 1 is not satisfactory to produce an absolute scale of measurement. At least 2 markers and 10 items were needed to attain a dependability index of 0.93; also 5 markers and 15 items were needed to attain a dependability index of 0.96 to successfully separate students in terms of performance irrespective of the performance of other students. The result is similar to [6] in which more invigilators were needed to attain high dependability index. The findings of this study found support also in the study of [21] which showed that an increase in the level of raters yielded a higher dependability index than when the raters were few.

Finally, it will take more than 1 marker to attain a high dependability index of 0.93 for absolute measurement that considers the object of interest in many educational assessments.

REFERENCES

[1] Johnson, S., Dulany, C. & Banks, K. (2000). Measurement error. Retrieved April 3, 2021 from http://www.wcpss.net/evaluation/research/reports/2000/memento.errx.pdf.

[2] Hofman, D (2005). Common sources of errors in measurement systems. Retrieved January 3, 2021 from http://wu.wiley.comlegacy/wiley enrich/hbuds/pdfs/mm154.pdf.

[3] Miller, M.D. & Linn, R.L. (2000). Validation of performance-based assessments. Applied Psychological Measurement, 34, 367-378.

[4] Thompson, B. (1994). Guidelines for authors. Educational and Psychological Measurement, 54, 837-847.

[5] Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. Educational and Psychological Measurement, 58, 6-20.

[6] Egbulu, C.A. (2013). “Estimating measurement error and score dependability in examinations using generalizability theory”. (Unpublished doctoral dissertation). University of Nigeria, Nsukka.

[7] Vanleeuwen, D.M. (1997). Assessing reliability of measurements with generalizability reliability. Journal of Agricultural Education, 38, 36-42.

[8] Bichi, A. A, Suleiman, A. H & Ali, H. (2019) Students’ achievement in Mathematics: Analyzing the influence of Gender and School Nature. Contemporary Educational Researches journal, 9(3.), 50-56.

[9] Anneh, V.G. (2014). WAEC releases Nov/Dec 2014 results, says 70% of students failed. Daily Post Nigeria. Retrieved from http://www.google.com/amp/s/dailypost.ng/2014/12/19/waec-releases-novdec-2014-results-says-70-students-failed%3famp.

[10] National Bureau of Statistics. (2019). WAEC Results Statistics (2016-2018). Retrieved from https://www.education.gov.ng/wp-content/uploads/2019/10/WAEC_RESULTS_2016-2018.pdf.

[11] Cardinet, Johnson & Pinion (2010). Applying Generalizability Theory using Edug. Quantitative Methodology Series, New York, Routledge.

[12] American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). Standards for Educational and Psychological Testing. Washington, D.C: American Educational Research Association.

[13] Johnson, S. & Johnson, R. (2009). Conceptualizing and interpreting reliability, covalidity: Ofqual. Shavelson, R.J., Baxter, G & Gao, X. (1993). Sampling variability of performance assessments. Journal of Educational Measurement, 30, 215-232.

[14] Shavelson, R. & Webb, N. (1991). “Generalizability theory; 1973-1980; British Journal of Mathematical and Statistical Psychology, 34, 133-166.

[15] Mahmoud, J. A. (2017), Analysis of Dependability of Undergraduate Students’ Scores in Teaching Practice Courses in a Nigerian University. An Unpublished PhD Thesis, University of Boran.

[16] Demorest, M.E., & Cord, M. (1993). Evaluation of temporal and interlist sources of variability in NU Test Scores: A generalizability analysis.

[17] Huang, J. “How accurate are ESL students’ holistic writing scores on large- scale assessments? A generalizability theory approach.” 2008. Retrieved from http://www.niagara.edu/assets/assets/nctelestdard33/33/6.DnHuang.

[18] Brennan, R.L. (2001). An essay on the history and future reliability from the perspective of replications. Journal of Educational Measurement, 38, 295-317.

[19] Webb, N.M., Shavelson, R.J., & Hartel, E.H. (2007). “Reliability coefficient and generalizability theory” Handbooks of statistics 26: Psychometrics (C. Rao and S. Sinharay Eds.). 81-124, the Netherlands: Elsevier.

[20] Lee, Y. (2006). “Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks, language Testing”, 23, 131-166.

Nkechi Patricia-Marie Esomonu is a Professor of Educational Evaluation, Research and Statistics with background in Science Education, Biology. She is a Fellow of Science Teachers’ Association of Nigeria (FSTAN). She is a member of Nigerian Academy of Education (MNAE), International Association for Educational Assessment (IAEA), World Council for Curriculum and Instruction (WCCI), Association of Educational Researchers and Evaluators of Nigeria (ASSENER) and Educational Assessment and Research Network in Africa (EANNA). She has been Editor of many volumes of Journal and have published many articles in peer review journals locally and internationally.

Prof P M Esomonu has served Nnamdi Azikiwe University in various capacities. She was the Director, Affiliate Institution Affairs, 2019; Dean, Faculty of Education -2016- 2019; Member, 7th University Governing Council, Nnamdi Azikiwe University, Awka, 2015 -2017, Associate Dean, School of Postgraduate Studies, Nnamdi Azikiwe University, Awka- 2015-
James Uzomha Okeaba is a Ph.D student of Nnamdi Azikwe University, Awka. He holds a Masters’ degree in Educational Measurement and Evaluation, a Bachelor’s degree in Mathematics Education. He is a Leadership Coach with a Basic and Advanced certificate in Leadership from the highly esteemed “Daystar Leadership Academy”. He is the author of the sensational book, “I am responsible” and the Lead Pastor at “The Living Souls Worship Assembly”.2017; Head of Department, Education Foundations, Nnamdi University among others.