International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 4, No. 02; Mar-Apr 2021

ESTIMATING THE EFFECTS OF THE INSTRUMENTATION FACETS AND THEIR INTERACTIONS ON SCORE DEPENDABILITY IN EXAMINATIONS USING GENERALIZABILITY THEORY.

Nkechi Patricia-Mary Esomonu and James Uzomba Okeaba
Department of Educational foundations, Nnamdi Azikiwe University, Awka, Ananmabara State, Nigeria.

DOI: http://dx.doi.org/10.37500/IJESSR.2021.4226

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 mathematics examinations, it is needful to estimate the effects of the instrumentation facets and their interaction on score dependability using generalizability theory. Three research questions and two hypotheses were posed to guide the study. The study population comprised 5,085 SS3 students of the 34 Government-owned senior secondary schools in Yenagoa LGA of Bayelsa State. Section A of the 2019 NECO Mathematics main paper and 2019 NECO Mathematics Marking Scheme were used to collect the data. EduG version 6.0-e based on ANOVA and Generalizability theory was used to answer the threes research questions. A 95% confidence interval was computed using the S E variance components to determine whether there was a significant difference in the instrumentation facets' effects and their interactions to measurement error and score dependability in examinations. The study's findings revealed that the facet items contributed about 18.5% of the measurement error while the facet makers had 0% effect on the score dependability. Items and markers were not significantly different in their contributions to score dependability and an index of dependability of 0.93 that is high enough to maximize reliability was obtained when we have level of markers at 2 and the items at 10.

KEYWORDS: Generalizability theory, instrumentation facets, dependability, students, Yenagoa,

INTRODUCTION

Examinations are often used to place students in the classroom or know how much they have learned in a given body of knowledge and for certification. Scores obtained from these evaluations are used to judge the students. The challenge with the scores obtained in the examinations is that there is degree of errors that affect these scores so that the score obtained by students in any exam is not a true representation of the students' ability. According to Egbulefu (2013), Test scores are not a definitive measure of student's knowledge or skills. An examinee's score can be expected to vary across different versions of a test. The score variance is often because of differences in the way the markers evaluate 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, no single test score can be a perfectly dependable indicator of a student's performance. Measurement (random) errors can result from factors such as the way the test is designed, students' individuality, testing situation or other sources such as examiner's mood, test time (occasion), test environment, invigilators, and the changing order of the questions, which may lead to higher or lower scores (Johnson, Dulany & Banks 2000). Some test items (questions) may be biased in favour of or against particular groups of students; the need for estimating measurement error arises because of the inconsistencies in measurements and the rate at which students fail.

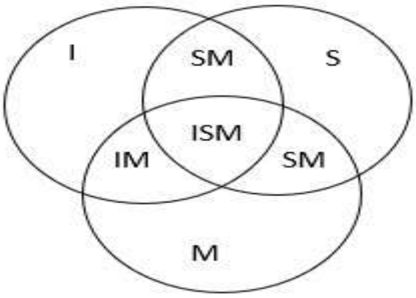
The West African Examinations council results in 2018 reflected that a total of 1.57m candidates sat for WAEC as public students. The results show that 48.15% had 5 credits and above, including English and Mathematics, while 51.85% failed to do so. In the same year, a total of 109,798 candidates sat for WAEC as private students, but only 33.81% had 5 credits and above, including English and Mathematics, while 66.19% did not (National Bureau of Statistics, 2019). The question here is; do these scores reflect the performance of students in the examination? This low performance of students in examinations calls for the estimation of multiple sources of error, in order 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 in examination scores.

Observed scores in examinations are affected by factors other than the student's cognitive ability. Specific factors (facets) such as test questions, invigilators, and markers are likely to affect the reliability of an observed score in examinations and affect interpretation and decision-making after that. The impact of these factors leads to questions about the accuracy, precision, and ultimately, the fairness of students' scores in examinations.

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, markers, Questions, Procedures, Conditions, Rules for scoring, etc. 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.

MATERIALS AND METHODS

Study design: The study's design was a random effect design, two-facets fully crossed s x I x m design for a generalizability (G) and decision (D)studies. The researcher used a fully crossed design in the G-study 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, markers (m)and the test questions items (I). Circle-overlap areas represent facet interactions, and the seven distinct areas correspond to the seven effects.



Population and Sample: The study population 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 because they are the only class ready for the NECO External Examinations and are 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 formed the sample for the study.

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

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

Data Analysis: EduG version 6.0-e which is based on the Analysis of Variance (ANOVA) and Generalizability Theory was used to carry out the Generalizability analysis. It will be used to answer the 4 research questions. To test the two hypotheses at 5% significant level using standard error variance components will be 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 will be 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.

RESULTS

TABLE 1: A G study showing the effects of questions, markers and their interactions to score dependability in examinations

SOURCE	VC ESTIMATE	RELATIVE ERROR VARIANCE	% RELATIVE	ABSOLUTE ERROR VARIANCE	% ABSOLUTE
STUDENTS (S)	4.13352				
ITEMS(I)	0.48160			0.09632	18.5
MARKERS(M)	-0.00201			(0.00000)	0.0
S x I	1.39004	0.27801	65.8	0.27801	53.4
S x M	0.10630	0.035443	8.4	0.03543	6.8
I x M	0.03133			0.00209	0.4
S x I x M	1.6369	0.10913	25.8	0.10913	20.9
TOTAL		0.42257	100%	0.52098	100%

Coefficients: EP^2 0.91; Φ 0.89

From the table 1 above, the facet items produced an absolute error variance of 0-09632 accounting for 18.5% of the absolute variance while the absolute error variance of the facet makers is 0.000 accounting for 0% of the absolute error variance.

The interaction of students and items yielded an absolute error variance of 0.27801 which accounts for 53.4% of the absolute variance also the interaction of students and makers produced an absolute error variance of 0.03543 which is 6.8% of the absolute variance. The interaction of items and makers produced 0.00209absolute variance which accounts for 0.4%of the absolute variance. Lastly, the residual which is the interaction of the three facets students x items x makers produced 0.10913 absolute error variance accounting for20.9%of the absolute variance.

TABLE 2: Estimated dependability index (ϕ) for a fully crossed s x i x m d-study design with different markers

LEVEL OF MARKERS	LEVEL OF ITEMS	Φ
1	5	0.84
2	5	0.87
3	5	0.89
2	10	0.93
1	10	0.90
5	15	0.96

Table 2 showed that with 1 marker and 5 items the dependability index (Φ) was 0.84, which has crossed the benchmark of 0.8. 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 items produced an increase of 0.09 (0.84 to 0.93) in the 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

Test of hypothesis 1

TABLE 3: 95% confidence interval on G-study Variance Components

VARIANCE COMPONENT	LOWER LIMITS	UPPER LIMITS
σ^2S	3.13313	5.13315
σ^2I	-0.519	1.4809
σ^2M	-1.002311	0.9977
σ^2SI	0.39	2.39
σ^2SM	-0.8937	1.1063
σ^2IM	-0.969	1.0313
σ^2SIM	0.6368	2.6368

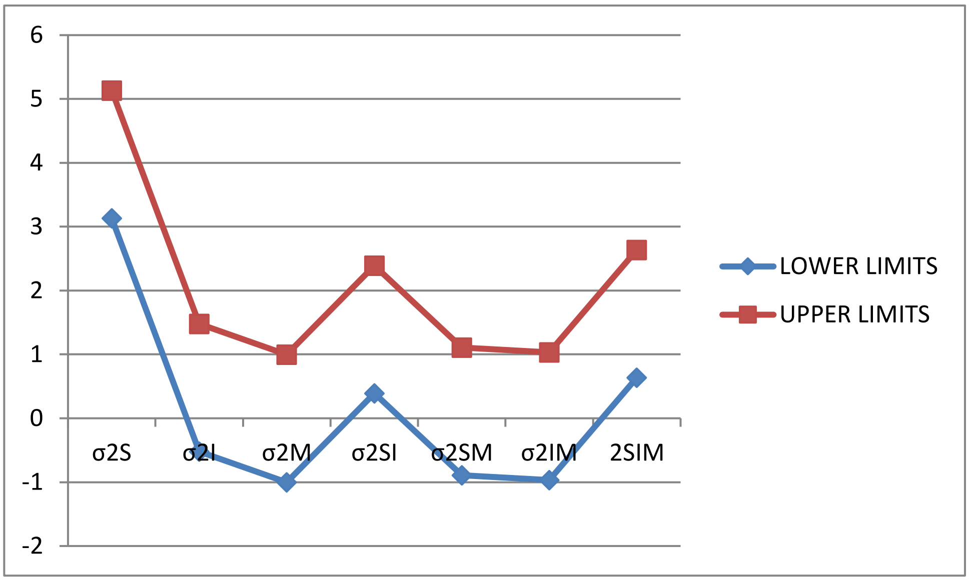


FIGURE 2: GRAPH INDICATING THE OVERLAP VARIANCE COMPONENTS

From table 3 and figure 2, the variance components of the instrumentation facets questions items σ2I and makers σ2m overlapped indicating that they were not significantly different in their effect on score dependability in examinations. Therefore, the null hypothesis is accepted.

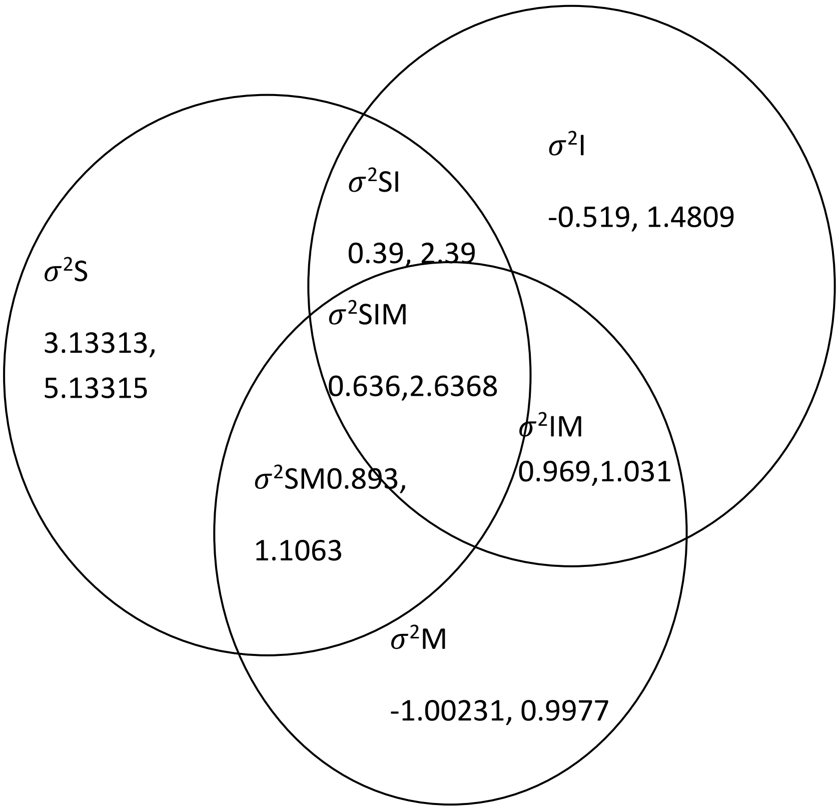


Figure 3: Venn diagram showing the overlap variance components of the facets.

Test of hypothesis 2: From table 3 and figure 2, it is clear that the interactions of the instrumentation facets σ^2si ,σ^2sm , σ^2im, and σ^2sim did overlap. The implication of this is that they do not differ significantly in their effect on score dependability in examinations. Hence the null hypothesis is accepted.

DISCUSSIONS OF FINDINGS

The study showed that the instrumentation facet item contributed 18.5% to measurement error and thus affecting score dependability while the facet makers did not contribute to measurement error and has little or no effect on score dependability. Hence from the D-study, increase in the level of items increased the score dependability. This is contrary to Egbulefu (2013) in whose result questions and interaction of students x questions did not have any effect on score dependability in examinations, in his study; increase in invigilators increased score dependability.

For the hypothesis, items and markers were not significantly different in their contributions to score dependability. This partly agrees with the findings of Bilger, Neutzel, Rabinowitz & Rzeezkowski (1984), in which the facet examiner was not statistically significant. However, in the same study, items are found to be statistically significant.

Students and items' interaction had the highest effect on score dependability producing an absolute variance of 0.27801 which is 53.4% of the absolute variance. This is followed by the residual σ2SIM Contributing 20.9% of the absolute variance. The implication of this result shows that students faced difficulties in answering the distributed items. This is in agreement with Demorest & Cord (1993), Huang (2008) and Egbulefu (2013) who noted that the interactions of students' x items ought to be the most important contributors to measurement error in an educational context. Also, the fact that the residual σ2SIM contributed greatly to error variance implies that other hidden factors were contributing to measurement error which is line with the findings of Shavelson, Baxter and Gao (1993) who also reported the residual effect as a major source of measurement error. This finding also found support in Shavelson & Webb (1991), whose study found the residual as one of the major contributors to measurement error. This study's result was also supported by the findings of Egbulefu (2013) who reported that the residual also made the highest contribution to measurement error. The study of Mahmud (2017) 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 (AERA, 1999).

The study further revealed that a dependability index of 0.93, which is exceptionally high enough to comfortably separate students who passed from those who failed was achieved when the level of items was 10 and the level of makers 2. When the level of markers was 1 and items 5, a high dependability index of 0.84 was produced, however the result shows that this index can be greatly improved to 0.93 by increasing the markers to 2 and items to 10. This can further be improved 0.96, if we increase the items to 15 and maker 5. These dependability indices are high enough to successfully separate students in terms of performance irrespective of other students' performance. The result is similar to Egbulefu (2013) in which more invigilators were needed to attain high dependability index. The findings of this study found support also in the study of Lee (2005) which showed that an increase in the level of raters yielded a higher dependability index than when the raters were few.

CONCLUSION

The result revealed that: The facet items contributed about 18.5% of the measurement error while the facet makers had 0% effect on the score dependability. Items and markers were not significantly different in their contributions to score dependability. (p>0.05). The study further showed that the interactions of the instrumentation facets were not significantly different in their contribution to score dependability and an index of dependability high enough to maximize reliability was obtained when we have level of markers at 2 and the level of items at 10.

REFERENCES

Bilger, R. C., Nuetzel, J., Rabinowitz, W.M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. Journal of Speech and Hearing Research, 27, 32-48.

Brennan, R.L. (1983). Elements of generalizability theory. Iowa City, Iowa: ACT Publications.

Brennan, R.L. (1984). Estimating the dependability of scores. In K.A. Berk (ed.), A guide to criterion-referenced test construction 292-334_. Battimore: Johns Hopkins University Press.

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

Brennan, R.L. (2003). Coefficient and indices in generalizability theory, Retrieved October 4, 2019 from http://www.education.uiowa.edu/casma/ASA.casma.rpt.pdf.

Brennan, R.L. (1991). Elements of generalizability theory. (2nd ed.). Iowa City, IA: The American College Testing Program

Brennan, R.L., Gao, X. & Colton, D.A. (1995). Generalizability analysis of work keys listening and writing tests. Educational and Psychological Measurement, 55 ,157-176.

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

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

Gao, X., Shavelson, R.J. & Baxter, G.P. (1994). Generalizability of large-scale performance assessments in science: Promises and problems. Applied Measurement in Education, 7, 323-342.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large- scale assessments? A generalizability theory approach. Retrieved October 6, 2019 from http:/www.niagara.edu/assets/assets/ncetelstadards/3/3.6f.DnHuang.

Hofman, D (2005). Common sources of errors in measurement systems. Retrieved January 3, 2010 from http://wu.wiley.com/legacy/wileychi/^''hbusd/pdfs/mm154 pdf.

Johnson, S., Dulany, C. & Banks, K. (2000). Measurement error. Retrieved April 3, 2009 from http://www.wcpss.net/evaluation.research/reports/2000/mment_error.pdf.

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

Thompson, B. (1991). Review of the book Generalizability theory: A nume Educational and Psychological Measurement, 51, 1069-1075.

Thompson, B. (1992). Two and one half decades of leadership in measurement and evaluation. Journal of Counseling and Development, 70, 434-448.

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

Thompson, B. &Vacha-Haase, T. (2000). Psychometrics is data metrics: The test is not reliable. Educational and Psychological Measurement, 60, 174-195.