Generalizability theory analysis of reliability of scores assigned to students in essay mathematics examinations with marking scheme in Akwa Ibom State

Prof. Mfonobong E. Umobong & Udeme E. Tommy
Department of Educational Foundations
University of Uyo, Uyo, Nigeria

Abstract

This study examined reliability of scores assigned to students in essay mathematics examinations with marking scheme based on generalizability theory in Akwa Ibom State. Two research questions were raised and a two-facet crossed design of the generalizability theory is used. The sample size for this study comprised 100 Senior Secondary Ill (SS Il) students and two mathematics teachers (raters) Mho were randomly sampled in Abak Metropolis of Akwa Ibom State, Nigeria. Mathematics

Essay Promotion Examination developed and administered by the Akwa Ibom State Ministry of Education for SS Il students with inter-rater reliability index was 0.72 was adapted and used as instrument for data collection. The collected data was analyzed using general linear model and the results obtained revealed that the generalizability and decision study variance components and interactions varied among students, items and raters based on their contribution to the analysis. The generalizability and dependability coefficient were found to be 0.69 and 0.69 respectively whichindicated high reliabilities. It was concluded that in adopting generalizability theory analysis, desirable reliability estimates could be achieved for scoring essay examinations. It was recommended that instead of using a single major essay examination during a term, schools should divide it into several smaller examinations and make use of multiple raters to grade each student response as these approaches have the advantage of improving reliability of scores.

Keywords: Generalizability theory, reliability, generalizability coefficient, dependability coefficient

Introduction

Mathematics is the bedrock of science, technology and other areas of study like economics, finance and accounting. It has the capability of producing advancement and evolution for any country and the world at large. The proficiency in Mathematics learning is extremely important to students and nations. Its efficacy is of great benefit in solving practical life problems, daily transaction dealings, discoveries and inventions, automation and robotics breakthrough. Good knowledge of Mathematics gives students great advantage in academics and this knowledge may be better measured by essay examinations (Williams et al., 2011). Essay examinations are essential component of educational assessment as they require high level of intelligence that needs students to broaden their thoughts, have good analytical skills, critically distinguish and justify facts in order to provide the answers required from them by the items in such examinations. However, a study (Arum & Roksa, 2011) reported that despite the fact that essay examinations are critical aspect of assessment, there is an alarming low level of students essays in schools nowadays. Moreover, the authors found that many students produce little academic demand in essay examinations. Essay examinations play crucial role in the development of critical thinking and analytical skills along with the ability to effectively communicate one's ideas to others. It is imperative that schools provide as many writing opportunities (essays) as possible to foster their development (Reed & Burton, 2015; Pare & Joordens, 2018).

The use of essay examinations in a course is, however, beset by several challenges. In addition to being resource and labour intensive for raters, grading of essay examinations is plagued by subjectivity and uneven variability (Anatol & Hariharan, 2009; Bell, 2010). The variations in the scores assigned by the same instructor or different instructors to the same paper may be the result of several factors. According to Coffman (2011), there are three categories of explanations for the variation in scores assigned to students' essay examinations. First, instructors may employ different standards in their ratings, with some being more lenient or severe than others. Second, some instructors distribute their scores over a greater portion of the rating scale, whereas others tend to concentrate their scores around a specific value. And third, instructors may differ in the criteria employed for rating the papers. Hence, if criteria are not pre-specified in the form of a marking scheme, scores may vary even if the same instructor grades the paper twice.

The relative difference in the level of difficulty of essay items may also contribute to the variation among different instructor ratings, while also compromising the impartiality of the scoring process (Barrett, 2009). Moreover, other scholars found that factors such as the student's first name and gender, the presentation of the answers, the language used in the essay and the order of the paper in the pile of essays to be marked may also influence the instructors' judgments (Brown, 2010; Branthwaite et al., 2011). The culmination of these problems is that the reliability of scores assigned to essay examinations is often very low (Hopkins, 2018).

Studies investigating the reliability of scores assigned to students in essay examinations using inter-rater reliability abound. A study by Haltog et al. (2016) reached a conclusion and reported that similarities of ratings among five instructors rating in essays varied between -0.41 and 0.85 with an average of 0.44. An even more alarming conclusion was drawn by Bull (2016) who reported that the grading of final year examination was so unreliable that a random assignment of grades could have been helpful in differentiating among the examinees. Blok (2015) investigated reliability of scores marked by 16 raters who separately graded 105 essay examinations in two separate situations using scale from I (very poor) to 10 (excellent). The study found that the estimated correlations among the scores of different raters ranged between 0.415 and 0.910, indicating that a significant variability existed in the rank-order of the scores assigned by different raters to the same papers. Fair levels of inter-rater agreement were also reported in a study that employed data from 13 examiners and 233 answer papers (k =0.385) (Anatol & Hariharan, 2009). Similarly, the overall reliability based on Cronbach's alpha coefficient was 0.672.

Given the problems reported in the literature, Coffman (2011) recommended the use of two raters to rate the same essay so as to improve the reliability of scoring. Cannings et al. (2005) made a similar recommendation based on the results of two study cohorts (1990-2000 and 2002-2003) which found reliability of grades of students' answers to be 0.38 and 0.39 respectively. Additionally, weighted Cohen Kappa measure of agreements among examiners' ratings produced a coefficient of0.42 between the examiners of the first cohort and 0.62 between the examiners of the second cohort. In contrast, Frijns et al. (2010) found a generalizability coefficient of 0.80 for open ended responses marked by physicians as the raters, where two raters received between four and six hours of training as reported by Kuper, (2006).

One way in which the reliability of essay examinations could be improved is through the use of marking schemes. In addition to saving time in providing feedback (Barringer, 2008), marking schemes explain different dimensions of question, tell raters concerning level of acquisition needed for a particular question, and illustrate condition on which they are to be scored (Hamer & Hafner, 2013; Stevens & Levi, 2015). Furthermore, by describing grading requirement in advance, marking schemes may significantly impact inter-rater reliability (Moskal & Leydens, 2010). However, even when marking schemes are employed, the reliability of scoring may not necessarily be very high. For example, Williams et al. (2011) investigated the inter-rater reliability of scores obtained by tutors in rating students' final papers. The eight raters were provided rubrics to assist them in the grading process, and required to grade all papers using a 12-point scale. The inter-rater reliability of their scores was 0.79, with a 95% confidence interval of0.49 to 0.93. If a rater marked a student's response 9 (B+) based on the 12-point scale, scores of other raters for the same student ranged from 6 (C+) to 12 (A+) 95%.

Given the difficulty in producing highly reliable scores in the scoring of essay examinations, it is not surprising that schools indicated more concern about the marking of students' essay examinations and felt that the scores assigned most at times were not reliable and mostly not accurate (Orpen, 2010). However, essay examination is a necessary aspect of assessment. Therefore, it becomes important for schools, teachers and administrators to find ways to reduce or eliminate errors in the scoring of essay examinations so as to increase the likelihood of producing grades with high reliability. Hence, essay examinations require that the scores assigned to them are reliable for accurate interpretations and decisions to be made about them. However, research on the reliability of the scores assigned to students in essay examinations reveal a high degree of contradiction, with some researchers concluding that the scores assigned are very reliable while others suggest that they are so unreliable and that random assignment of scores would have been almost helpful. Educational researchers have for long dealt with the issue of reliability in different ways. Thus, generalizability theory (GT) is one of the approaches propounded for assessment of reliability of scores rather than using coefficient alpha since GT considers the different sources of errors of measurement not addressed in classical test theory (CT T) frameworks.

In CTT, there is a linear model that shows that observed test score (X) is the sum of true score (T) and error score (E). GT takes care of disadvantages of CTT as it distinguishes the multiple sources of error using analysis of variance methods (Briggs & Wilson, 2007). In GT terms the objective of measurement is to measure the qualities of subjects and facets are likely sources of measurement error except object of measurement. For instance, in a mathematics essay test, ability of students represents object of measurement while the items, the rater(s) and test form are facets. Generally, expected grades of respondents are different from observed grades. Expected grades are gotten from all available facets while observed grades are gotten from sampled facets. The variance between expected grade and observed grade can partially be gotten from facet-based measurement errors.

GT gives a vivid illustrative conceptual framework and strong set of statistical processes for resolving many measurement problems such as reliability. Even though statistical dimension of GT are very useful, in fact, the major outstanding quality of GT is its conceptual framework that allows a multifaceted perspective on measurement error and its components. Furthermore, GT enables one to estimate the number levels (sample size)

necessary' for each facet in order to attain a desired reliability level. GT framework has two aspects, generalizability (G) study and decision (D) study. G study segregates variance components Into multiple sources of error. D study quantifies universe score variance, error variances and measurement precision coefficients based on the G study (Brennan, 2001). Paramount contribution of GT is that it allows a decision maker to ascertain the sources of measurement error and use appropriate number of observations correctly so as to

achieve a required amount of generalizability (Marcoulides, 2013). Sources of measurement error are estimated in G-study. Decisions are then taken on each of the sources, that is, which is small enough to be overlooked or, better still, which error sources allow decrease in number of relevant observations in subsequent decision study without greatly decreasing the generalizability coefficient (that is, reliability). Since resources are very scarce, such information could be of help in justifying suitable means using them. Based on these two purposes, estimation of variance components in a G study design is unarguably the main aim. In Akwa Ibom State secondary school system, it has been observed that the state minist1Y of education sets essay mathematics examinations with very few items that teachers can finish marking as soon as possible without wanting to know if the examinations have enough content coverage or not. It administers, marks the examination and produce scores for students involved and nobody cares whether the scores are reliable or not since there is no comparison of such scores with that of other teacher (s). Also, there is no averaging of scores to obtain a single score since the scores are assigned by just a single teacher. It is as a result of these issues that researchers have suggested that teachers and schools ought to use many raters to score students' essay responses to essay tests. They also suggest schools and teachers to give many writing opportunities to students in the form of essay examinations to foster their intellectual development and the use of two or more raters to rate the same answers to essay examinations as it may lead to good reliability of scores. It is against these concerns that generalizability theory analysis of the reliability of scores assigned to students in essay examinations was carried out with the objectives of providing explanations, interpretations of coefficients and indices propounded for usage in GT such as variance components, generalizability and dependability coefficients as the coefficients are extremely important if they are interpreted accurately.

Objective of the Study

This research aimed at carrying out a theory analysis of reliability of scores assigned to students in essay mathematics examinations with marking scheme in Akwa Ibom State but the objectives of this research were:

l. To assess the valiance components of students, raters and items in the generalizability theory analysis.
2. To examine the generalizability and dependability coefficients in the generalizability theory analysis.

Research Questions

l. What are the variance components of students, raters and items in the generalizability theory analysis?
2. What are the generalizability and dependability coefficients in the generalizability theory analysis?

Methodology

To differentiate variance components, GT uses experimental design procedure and based on relationship among possible facets, a two-facet crossed design (pxixr) of GT was used. In this design, p stands for object of measurement which is the students, i represents the test items and r represents the raters which means that every student need to respond to every items which are marked by all raters. In GT Fundamental assumption, the persons (p), items (i) and raters (r) are sampled independently and randomly from population of persons, items and raters. The participants for this study were 100 senior secondary three students who were randomly sampled from a population of 480 senior secondary three students in the 10 public secondary schools through simple random sampling technique in Abak Metropolis in Akwa Ibom State, Nigeria. TWO mathematics teachers who served as raters in this study were also randomly sampled from twelve mathematics teachers who teach senior secondary classes in the study area using simple random sampling technique also. The students' mean age was 16.5 and they were almost equally divided by gender, with 53.7% of students being female and 46.3% male. Mathematics Essay Promotion Examination developed and administered by the Akwa Ibom State Ministry of Education for SS Il students in 2018 was adapted and used as instrument for data collection. The essay mathematics examination comprised 4 essay items on surds, matrices and determinants, logarithm and arithmetic of finance. It was validated by three mathematics teachers who have taught mathematics in senior secondary three classes for more than ten years. To estimate the reliability index of the instrument, it was administered on thirty senior secondary three students and their responses were rated by two raters (mathematics teachers). The grades of the two raters were subjected to inter-rater reliability analysis and the reliability index was estimated to be .72 which indicated that the instrument was reliable for use in carrying out the study. A marking scheme developed by the ministry of education to guide the raters in assigning grades to the students' responses was used. The marking scheme broke down the tasks pertaining to each item into objective criteria. All papers were graded on a 50-point scale with each item receiving a total score of 12.5. Two raters (mathematics teachers) were sampled to grade all the papers and were given the descriptions of each item and their respective scores. No additional training was provided for the raters. The collected data were analyzed using General Linear Model and Minimum Non Quadratic Estimation.

Data Analysis and Results

The generalizability analysis consisted of two parts, G-study and D-study. G-study is similar to a pilot study that utilizes a specific study design
(example, fully crossed) and is conducted under a set of conditions, called universe of admissible observations, defined by the researcher(s) based on assumption of fixed, random or mixed model variables. The D-study represents the study design and conditions known as the universe of generalization (that is, population and conditions that the researcher wants to generalize the results to) under which the study was conducted in the future. Based on these conditions and the variance estimates obtained in the G-study, the researcher can compute a generalizability (reliability) coefficient as well as dependability coefficient.

Research Question I

What are the variance components of students, raters and items in the generalizability theory analysis?

In order to answer research question l, the variance components of students, raters and items in the generalizability theory analysis is presented in Table I.

Table I: Variance components of students, raters and items in the generalizability theory analysis

Sources of variance		df	MS	GVAR		Sources of variance	DVAR	
students (p)	23484.91	99	237.22	25.17	26.62		25.17	68.94
items (i)	432.38	3	144.13	0.43	0.45	1	0.11	0.30
raters (r)	1.65	1	1.65	0.05	0.05		0.03	0.08
	23843.86	297	80.28	13.71	14.50	PI	3.43	9.39
	4392.92	99	44.37	2.13	2.25	PR	1.13	3.11
Ir	92.68	3	30.89	0.22	0.23	IR	0.03	0.08
1)11', e	15649.59	296	52.87	52.87	55.90	PIR	6.61	18.10

Students 100, Items = 4, Raters = 2, GVAR=G study variance, DVAR D study variance

Table I shows the analysis of variance results of the generalizability analysis. An examination of the G study and D study variances (0.05 and 0.03 respectively) in the table revealed no significant variability in the scores assigned to the students by the raters across the four items. G study and D study show that the variance components for students, items and raters (25.17, 0.03, 0.05 and 25.17, 0.11 and 0.03 respectively) and their interactions (13.71, 2.13, 0.22, 52.87 and 3.43, 1.13, 0.03, 6.61 respectively) varied greatly. There are two large variance components, that of

interaction plus error which is the highest (52.87 and 18.10 respectively) and then that of the students (25.17 and 25.17 respectively). That of interaction might have been filled with some uncontrolled sources of variance in which little

is In own and this ought to be considered in subsequent studies. The high student variance component reveals that the students are different in their responses to the items based on their mathematics abilities. Similarly, the high student-item interaction variance components (13.71 and 3.43 respectively) indicate that the students' responses differ on the items. Therefore, students' performances differed by

items. Item variance components (0.43 and 0. I I respectively) have low percentage of explaining the total variance, which signifies that the four problems posed by each item were of the students' level. The variance components (0.05 and 0.03 respectively) from the raters have low percentage of explaining the total valiance, which is an indication that the agreement among raters was very high. Students x item common effects (13.71 and 3.43 respectively) and students x rater common effects (2.13 and I .13 respectively) show that raters assigned similar scores on the items. Thus, the raters' scores did

not differ by students which indicate the difference between the students regarding their performances. Thus, individual differences can be determined using the marking scheme. Item x rater common effects (0.22 and 0.03 respectively) show that raters did not score the items differently but gave reliable scores.

Research Question 2

What are the generalizability and dependability coefficients in the generalizability theory analysis?

In order to answer research questions 2, the generalizability and dependability coefficients are presented in Table 2.

Table 2: Error variances and reliability coefficients of the generalizability analysis

Model and error variances		Reliability coefficients	
02(ö)	11.17	Generalizability coefficient p2	0.69
G2(A)	11.34	Dependability (D	0.69
Universe score	25.17		

Table 2 presents error variances, universe score and reliability coefficients. The relative en•or variance denoted by 0^2 (ö) is used for normreferenced comparisons (that is, comparison of the score of a student with the scores of other students) while the absolute error variance denoted by 0^2 (A) is used for criterionreferenced comparisons (that is, comparison of the score of a student to a single fixed standard). The aforement oned variance estimates can be utilized in computing two reliability coefficients, the generalizability coefficient ? p^2 and the dependability coefficient 0. The generalizability and dependability coefficients were found to be 0.69 and 0.69 respectively which reveal that scoring with a marking scheme is more reliable.

The generalizability coefficient is the equivalent of the parallel test reliability used in CTT. That is, generalizability coefficient is equal to ratio off true score variance to observed score variance. The higher the G coefficient, the better the measurement procedure. The index of dependability is a measure of criterion reliability and denotes the probability that the absolute decision, resulting from a comparison of a student's score to a standard would replicate if the essay examinations were graded several times by a random set of raters under parallel conditions. Therefore, the index of dependability is a function of the location of the standard. The closer the standard is to the grand mean, the lower the index (likelihood) will be that the unknown universe score underlying the composite average of a randomly selected student would be correctly classified relative to the standard, and vice versa. The value of D coefficient is higher than that of G-coefficient. To increase its value, one of the facets of measurement will have to be increased. For this increase to be effective, the increment could be as high as times two the number of items and raters used in the G-theory analysis.

Discussion ofF indings

This study was a generalizabilitytheory analysis of the reliability of scores assigned to students in essay examinations. Great deal ofresearches has investigated reliability of scores of essay examinations. The results of these studies have been inconclusive with some studies reporting low levels of reliability (Anatol & Hariharan, 2009; Cannings et al., 2015), a few studies reporting good levels of reliability (Frijns et al., 2010; Williams et al., 2011) and yet other studies reporting mixed levels of reliability (Hartog et al., 2013; Blok, 2015). This uncertainty has led several researchers to argue against the use of essay test examination. The present study contributes to this body of knowledge by providing a plausible explanation for some of the contradictory results. Namely, most of the previous reliability studies employed analytical techniques that, by today's psychometric standards, are antiquated (example, kappa, correlations, Cronbach alpha). Generalizability theory, particularly the two-facet model, is arguably the most (or one of the most) sophisticated method(s) for estimating reliability. Therefore, the reliability coefficient reported herein is a "cleaner" measure of the reliability involved in grading students because it controlled for two sources of measurement error (items and raters).

This study demonstrated the grading of essay examinations can be very reliable (G-coefficient = 0.69 and D-coefficient = 0.69) even when only two raters were employed to rate only 4 items, provided that clear grading criteria are used (in the form of a marking scheme) and an appropriate study design is implemented. Emphatically, unlike the findings reported in previous studies, neither of the markers employed in the present research had prior training on how to use the marking scheme. Hence, it is conceivable that a slightly higher reliability coefficient could have been attained had both raters received such training. As expected, reliability was a function of both number of markers and number of items in the essay mathematics examinations. Increase in number of items and number of markers would produce higher levels of reliability. Therefore, instead of using a single major essay test during a term, teachers should divide it into several smaller examinations. This approach has the advantage of not only improving reliability, but it makes students put forward their ideas constructively and to see their progress from one examination to the next. Of course, teachers can use only a few numbers of items in essay examinations per term since they are time consuming to mark but the reliability would be very small when only few items are used in a terminal examination.

Conclusion

Essay examinations are important part of evaluation despite their weaknesses; they remain an important tool for assessing educational achievement. Therefore, it is imperative that means of increasing the reliability of scaling such assignments are found such as generalizability theory analysis. This study does not claim to have found the solution to the problem, but it successfully showed that in appropriate situations and by using the right study design, good reliability could be achieved for marking essay examinations. Furthermore, it illustrated the use of generalizability theory for estimating inter-rater reliability. Hence, it would behove on researchers not familiar with this analytical technique to explore its many benefits.

Recommendations

From the findings of the study, the following recommendations were outlined.

1. Instead of using a single major essay examination during a terms teachers should divide it into several smaller examinations as this approach has the advantage of not only improving reliability, but it gives students the opportunity to practice their and analytical skills multiple times during the term and to see their progress from one examination to the next.
2. Schools and teachers should use different teachers to mark the responses of students to essay examinations as it is possible to improve the reliability of the scores assigned to students in essay examinations when numerous raters are used.
3. Marking scheme should be carefully developed and used as a guide for marking essay examinations by teachers as it ensures high dependability of scores assigned to students by teachers.

References

Anatol, T., & Hariharan, S. (2009). Reliability of the evaluation of students' answers to essay- type questions. Indian Medical Journal, 58(1), 13-16.

Arum, R., &Roksa, J. (2011). Academically adrift: Limited learning on college campuses. University of Chicago Press.

Barrett, S. (1999). Question choice: Does makers variability make examinations a lottery? ^{N}MIat Do \hlue in Higher Education? Cornerstones, 12-15.

Barringer, S. A. (2008). The lazy professor's guide to grading: How to increase student learning while decreasing professor homework. Journal of Food Science Education, 7, 47-53.

Bell, R. C. (2010). Problems in improving the reliability of essay marks. Assessment &Evaluation in Higher Education, 5(3), 254-263.

Blok, H. (2015). Estimating the reliability, validity, and invalidity of esssay ratings. Hafner, J. C., & Hafner, P. M. (2003). Journal for Educational measurement, Quantitative analysis of the niblic as an 22(1), 41-52.

Branthwaite, A. , Trueman, M., &Berrisford, T. (2011). Unreliability of marking: Further evidence and a possible explanation. EducationReview, 33(1), 41-46.

Brennan, R. L. (2001). Generalizability theory. Springler-Verlag.

Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modelling. Journal of Educ ation a I Measurement, 44, 131—155.

Brown, G. T. (2010). The validity of examination essays in higher education: Issues and responses. Higher Education Quarterly, 64(3), 276-291.

Bull, G. M. (2016). An examination of the final exammation in medicine. The Lancet, 271,368-372.

Cannings, R., Hawthome, K., Hood, K., & Houston, H. (2005). Putting double marking to the test: A framework to assess if it is worth the trouble. Medical Education, 39(3), 299-308.

Coffman, W. E. (2011). On the reliability of ratings of essay examinations in English. Research in the Teaching of English, 5(1), 24-36.

Eelis, W. C. (1930). Reliability of Repeated Grading of Essay Type Examinations. Journal of Educational Psychology, 21 48-52.

Flijns, P., Van der Vleuten, C., Venvijnen, G., van Leeuwen, Y., &Wijnen, W. (2010). The effect of structure in scoring methods on the reproducibility of scores oftests using open-ended questions. In W Bender, R. Hiemstra, A. Scherpbier& R. Zwierstra (Ed.), Proceedings of the Third Interna tional Conference on Tea Ching andAssessing Clinical Competence (pp. 466—471). BoekWerk Publications.