APPLICATION OF GENERALIZABILITY THEORY IN ESTIMATING SCORE DEPENDABILITY OF ECONOMICS ESSAY TEST

FRANCIS ELOCHUKWU IKEH¹, MERCY IFUNANYA ANI², IJEOMA AWA KALU², ROSELINE
IFEOMA IKETEAKU³, BEATRICE A. EZE², BARNABAS C. MADU¹, CATHERINE U. ENE^{1*}

¹Department of Science Education, University of Nigeria, Nsukka, Enugu State, Nigeria

²Department of Educational Foundations, Alex Ekwueme Federal University, Ndufe Alike Ikwo, Ebonyi State, Nigeria

³Department of Educational Psychology, School of Education, Enugu State College of Education (Technical), Enugu, Enugu State, Nigeria

ABSTRACT

This study was aimed at estimating the score reliability of Economics essay test using Generalizability Theory, basically to: (a) estimate potential sources of error due to students (persons), items (tasks) and raters (judges) and (b) to obtain reliability estimates of the measurement instrument by increasing the levels of conditions in each facet. Variance due to student by item by rater interaction was the highest contribution of error variance followed by students' factor. It was also gathered from the analyses of the findings that it would be more efficient to increase the number of items rather than increasing the number of raters in order to maximize the score reliability of Economics essay test. Conclusions drawn from this study were analysis of scores from ratings of students in Economics essay test.

KEYWORDS: Generalizability Theory, Essay Test, Score Dependability, Economics

Received: Dec 10, 2020; **Accepted:** Dec 29, 2020; **Published:** Apr 15, 2021; **Paper Id.:** IJMPERDJUN202117

INTRODUCTION

Essay test is one of the assessment tools used in testing or assessing students' academic achievement in any given instruction, especially when teachers want the students to originate, organize, express, and integrate ideas in a given problem. Essay test, according to Onunkwo (2002), is a test in which students are required to provide answers to questions and therefore offers students the opportunity to organize and express their ideas in writing. Hence, it is particularly useful in evaluating objectives which deal with selection, arrangement, organization and expression of ideas. In the light of this, essay tests usually provoke critical thinking, originality and ingenuity (Gronlund, 1971).

Reiner, Bothell, Sudweeks, and Wood (2002) are of the view that essay tests challenge students to create responses rather than to simply select a response; that is, they usually require a student to compose an answer rather than selecting an answer. Onunkwo (2002) opined that teachers use essay tests because they have the potential to reveal students' abilities to reason, create, analyze, synthesize, and evaluate. This implies that it will enable the teacher to assess students' understanding of and ability to think about subject matter content. Reiner et al (2002) opined that essay type requires the students to demonstrate their reasoning and thinking skills, which give teachers the opportunity to detect problems students may have with their reasoning processes. When educators detect problems in students' thinking, they can help them overcome those problems. These above explanations of essay test indicate that essay tests can be used to: assess higher-order or critical thinking skills, evaluate students' thinking and reasoning skill; provide authentic experience, test thought process (Reiner et al, 2002), and offer students the opportunity to organize and express their ideas in writing (Onunkwo, 2002).

However, assessments that require examinees to provide extended constructed responses like essay test often lack score reliability (Cumming, Kantor, Powers, Santos, & Taylor, 2000; Miller & Linn, 2000) and such a test depends on subjective scoring due to the fact that students organize their responses to questions in different ways, and unfortunately allocating marks for such a test (essay-type) can be very unreliable (Baird, Greateorex, & Bell, 2004). When teachers assess students' achievement with the use of essay test according to Kim and Wilson (2009), the major problem of controlling the reliability of scores in different ratings are being faced. Various sources of systematic variability in measurement can contribute to errors in essay assessment, and thereby affecting the reliability of scores (Lin, 2014). According to Shavelson, Baxter and Gao (1993), students' achievement scores may vary across a sample of assessment tasks, raters, occasions, or methods.

Task and rater variability have been so far reported by researchers like Bachman, Lynch, and Mason (1995), Brennan, Gao and Colton (1995), Henning (1996), Lynch and McNamara (1998), Breland et al., (1999) and Fulcher (2003) as potential sources of measurement error in any assessment like Economics essay test. Measurement error according to Johnson, Dulany and Banks (2000) is a situation in which students' true ability or achievement is either underestimated or overestimated. Therefore, the need for estimating measurement error arises as a result of the inconsistencies that exist in measurements especially those involving multiple sources of error. Atilla (2012) opined that the use of classical test theory in determining score reliability, however, is not capable of identifying and untangling the profusion of error which classical reliability was not conceptualized to do. This, because it accounts for only one source of error at a time, as well posed a serious deadlock, since there are several potential errors which can exist simultaneously in any measurement. Other potential sources of error that remain are but undifferentiated error. Similarly, Tavako and Brennan (2013) opined that in the context of Classical Test Theory (CTT), students' true score is the sum of the students' observed score and a single undifferentiated error term. However, with the use of classical test theory, Cronbach's alpha has been the most frequently reported estimation of reliability, which includes only errors associated with sampling of items, that is, its estimate of reliability, is based on the covariation among items internal to the test (Tavako & Brennan, 2013). Meanwhile, the use of alpha (α) does not allow one to identify, locate, isolate, determine and equally estimate the impact of different sources of measurement error associated with observed student scores. Therefore, a more advanced method of Classical Test Theory known as Generalizability theory (G-Theory) is needed which is capable of isolating and as well as handling multiple source of error in a single analysis (Bergee, 2007).

The G-theory, therefore, is a statistical theory for evaluating the reliability of behavioral measurements which offers the researcher the opportunity to assess comprehensively various sources of measurement error that may arise at the same time (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). G-theory also concerns itself about the reliability or dependability of behavioural measures. Kim and Wilson (2009) refers to dependability of behavioural measures as the accuracy of generalizing from a person's observed score on a measure or a test to the score that the person who has received averaged over all possible conditions. G-theory according to Brennan, (2001) decomposes, estimates, and reveals all measurement errors, which are known as variance components and can be used to assess the reliability of aggregate-level measures of essay test (O'Brien, 1990). Shavelson and Webb (2005) stated that the use of G-theory allows researcher to address the following questions like: Is the sampling of items or raters the potential source of error in a measurement? Can reliability of measurement be better improved by increasing the number of items or raters, or is the combination of the two more are effective in improving score reliability? In certification decision, are the test scores

adequately reliable to make decisions about the performance of the students? Some researchers like Atılgan and Tezbaşaran (2005); Kan (2007); Yelboğa and Tavşancıl (2010) also agreed that the purpose of a G-study is to estimate multiple sources of error variance simultaneously like error variance due to persons (students), raters (scorer), and items (questions) within a single, multifaceted experiment. Generalizability theory provides a comprehensive structure for making value judgments about the multiple elements of observational systems as well as providing empirical evidence regarding the optimal number of raters (that is, whether to increase or decrease the number of raters) and the best practice of producing a desired reliabilities rather than grounding such choices in common practice (Healthier, Charalambos, & Mathew, 2012).

From the ongoing, therefore, it should be understood that in measuring students' performance in any given examination, there are other characteristics than the students' factor that may affect the reliability of an observed score in any given examination. Such likely factors also known as facets include the person, the item, and the rater, among others. There is a need, therefore, to estimate the measurement error and score reliability of such factors using generalizability theory and determine their contributions and their interaction effects in an observed score as well as identifying the differences in the reliability coefficient by increasing the number of conditions in each facet.

METHOD

The study employed two-facet fully crossed G-study and D-study designs [SxIxR]. Fully crossed design was used to estimate all possible variance components in the measurement situation, while the D study in generalizability theory uses the information obtained from the G-study to design the best possible measurement procedure in minimizing undesirable sources of measurement error and maximize reliability.

Economics Essay Test (EET) developed by the researchers was used to collect pertinent data for the study. The Economics Essay Test (EET) was based on the following Economics topics: Demand and Supply, concept of money, Agriculture, Distributive trade, and Production drawn from SSI&2 curriculums. Test blueprint or table of specification was used in preparing the test which was based on the six levels of cognitive domain of Bloom's taxonomy of education. The test contains five items (questions) with sub questions in each item, and each item carries an equal mark of (10) totaling (50) in the whole test. A well developed scoring guide by the researchers was used for scoring the instrument (Economics Essay Test) developed by the researchers.

To determine the face validity of the instrument, the instrument was given to two experts in Department of Social Science Education (Economics Education Unit) and two in Department of Science Education (Measurement and Evaluation Unit), Faculty of Education, University of Nigeria, Nsukka, Nigeria. Table of specification (test blueprint) was used to ascertain the content validity of the instrument. The reliability of the 5-item instrument was established using 22 SSII students of public secondary school which was an equivalent subject to the subject under investigation. The reliability coefficient of 0.93 was obtained using Kendal coefficient of concordance.

The instrument for the collection of relevant data were administered and collected by the researchers with the help of the research assistant. After the collection of the relevant data, the researchers assigned each student script to four (4) independent raters (Economics teachers) for scoring after they had been formally trained. During the training session, the validated instrument and scoring guide were discussed. There was also trial scoring of the dummies by the scorers during the training exercise to ensure mastering. The data collected was analyzed using error variance which was performed using

EduG version 6.0-e which was specifically developed for generalizability analysis which was used to answer the research questions.

RESULTS

Research Question One

What are the contributions of the facets; students, items and rater and their interactions to measurement error in Economics Essay Test?

Table 1: G-Study Analysis of Estimated Variance Components for Students, Item, Rater and Interactions among Facets

Source of Variance	Degree of Freedom	Sum of Square	Mean Square	Estimated Variance Component	% of Variance
Student	26	335.6037	12.9078	0.5518	15.7
Item	4	30.2667	7.5667	0.0263	0.7
Rater	3	3.3685	1.1228	0.0055	0.0
SI	104	398.9333	3.8359	0.4263	12.1
SR	78	145.9815	1.8716	0.3743	10.6
IR	12	27.6593	2.3049	0.0065	0.2
SIR	312	664.7407	2.1306	2.1306	60.6
Total	539	1606.5537			100%

Result in Table Iabove revealed the results of the estimated variance components and its proportions of the G-study. The table shows that the largest contribution to measurement error in the scores obtained comes from the residual σ^2_{sir} (2.1306) accounting for 60.6% of the total variability. This was followed by student factor σ^2_s (0.5518, 15.7%). The rater (σ^2_r) factor contributes a little and insignificant variance of 0.0055 accounting to 0.0% of the total variability. The estimated variance component of the interaction between the student and the item (σ^2_{si}) was 0.4263 which accounts for the 12.1% total variability. However, the variance obtained due to interaction between student and rater (σ^2_{sr}) was 0.3743 with total variability of 10.6%.

Research Question Two

What are the differences in reliability coefficient of Economics essay test by increasing the number of conditions (optimization) in each facet?

Summary of Estimated Reliability Index (Φ) for Fully Crossed (S x I x R) D-Study Results with Different Levels of Conditions in Each Facet (Raters and Items)

Facets	Initial Condition	Optimization				
Number of Raters	4	8	12	16	20	24
Reliability Estimate (Φ or Φ_i)	0.71	0.77	0.79	0.81	0.81	0.83
Number of Items	5	10	15	20	25	30
Reliability Estimate (Φ or Φ_i)	0.71	0.83	0.88	0.91	0.92	0.94

Results in Table 2 above show the differences in the reliability coefficient that result from changing or increasing the number of both items and raters. The increase in both levels of items and raters showed a steady and gradual increase. When the level of raters was increased from 4 to 8, a reliability index of 0.77 was obtained. Moreover, a more reliable reliability index of 0.83 was obtained from increasing further the number of raters from 4 to 24 thereby recording a

reliability index difference of 0.12. Secondly, the same steady and gradual increase was also noticed from increasing the number of items as well. A reliability index of 0.83 was recorded from increasing the items from 5 to 10 while a more reliability index of 0.94 was obtained by further increasing the number of items from 5 to 30 thereby recording a reliability index difference of 0.23.

DISCUSSION

The result of our findings as shown in Table 1 revealed that the largest contribution to measurement error in the scores obtained comes from the residual σ^2_{res} (2.1306) accounting for 60.6% of the total variability. The result indicated that a proportion of the variance was due to the interaction of students by item by rater (σ^2_{sir}) and other source of error variance that were not measured which indicates that students' performances were inconsistent across the sample of different tasks performed. It should be noted, however, that this large variance component recorded is not only related to items and raters but also to undifferentiated error.

The second largest source of error was the student (σ^2_s) factor with variance component of 0.5518 accounting for 15.7% relative to variation due to raters (σ^2_r) 0.0055 which accounts to 0.0% of the total variability indicating that error due to rater is minimal which indicates that the rater mean scores are very similar and this may be due to the level of details provided in the scoring guide and the nature of the training received during the training sessions for the raters. This is why Dunbar, Koretz, and Hoover (1991) have opined that, the effects of rater variance can be controlled with well specified scoring guide and appropriate training of raters. Moreover, it should be observed from the result that the generalizability of scoring substantially influenced by item difficulties. The estimated variance component of the interaction between the student and the item (σ^2_{si}) was 0.4263 which accounts for the 12.1% total variability, which indicates that even though the raters used the same marking guide, they also disagreed somewhat in the relative standard of students. This also shows that while the items were on average and comparable in difficulty, they were not uniformly difficult for all students that took the test. The findings of this study are in agreement with Shavelon & Webb (1991) who in their study found that residual as the largest contributor to error in measurement. The study also found that raters did contribute to a total score variation to some extent, but their overall effects on the Economics essay scores were relatively small, compared to those items by rater interaction and items that recorded little error variance.

In terms of the reliability of Economics, essay test obtained from increasing both number of items and raters, Table 2 showed a steady and gradual increase. When the level of raters was increased from 4 to 8, a reliability index of 0.77 was obtained. Moreover, a more reliable reliability index of 0.83 was obtained by increasing the number of raters from 4 to 24 thereby recording a reliability index difference of 0.12. Secondly, the same steady and gradual increase was also noticed from increasing the number of items as well. A reliability index of 0.83 was recorded from increasing the items from 5 to 10, while a more reliable reliability index of 0.94 was obtained from further increasing the number of items from 5 to 30 thereby recording a reliability index difference of 0.12. The minimal number of both items and raters was not quite satisfactory to produce an absolute scale of measurement. In the aspect of the items for example, 24 raters were needed at least to produce a reliability index of 0.94 to successfully separate students in terms of performance irrespective of the performance of other students. Though with the increase of both items and raters, there was a steady and gradual increase in reliability index, but, however, increasing the number of items produces a better reliability coefficient quickly than increasing the number of raters. The findings of this study are in agreement with the study of Lee (2005) who found that increasing the number of tasks had a relatively larger impact on the score dependability than when the number of

ratings was increased. The study also found support in the study carried out by Kim and Wilson (2009) in which the increase in the number of items produced a better generalizability coefficient more quickly than increasing the number of raters. Nevertheless, the findings of the study are also in consistent with the research studies on performance based assessment carried out by Gao, Shavelson and Baxter (1994) and Miller and Linn (2000).

CONCLUSIONS

On the basis of the results of the generalizability analysis, it was concluded that the largest contribution of error variance obtained in this study was from the residual factor, that is, student by item by rater factor followed by the students' factor. While the third largest variance component contributing to error variance was found to be associated with the student by item interaction followed by students by rater interaction. Secondly, an increase in the levels of both items and raters has a greater effect on increasing the generalizability ratings. But the increase in number of items (tasks), obtains a more optimal reliability coefficient than increasing the number of raters (judges).

Recommendation

The following recommendations are made based on the findings of this study so as to reduce the impact of the source of variability that may arise in scores obtained by students in any given examination and as well increase the score reliability in Economics essay test.

- Examination bodies should ensure that students' scores are subjected to generalizability analysis in order to estimate and reduce the various source of error variance to maximize reliability.
- In order to increase the score reliability (dependability) of Economics essay test, it is important to increase the number of items (that is, tasks) than increasing the number of raters (scorer).

ACKNOWLEDGEMENTS

The researchers appreciate all the research participants for their contributions, especially Dr. Catherine U Ene (Catherine.ene@unn.edu.ng) who served as the corresponding author for this research study.

REFERENCES

1. AERA, APA & NCME (1985). *Standards for educational and psychological testing*. Washington DC: American Psychological Association.
2. Ababneh, Ala'A. A., Sarah M Al-Ja'freh, and Lubna Abushaikha. "Traumatic Childbirth: Incidence, Risk Factors, and Its Impact on Mothers and Their Infants a Scoping Review." *International Journal of Applied and Natural Sciences (IJANS)* 6.6 (2017): 1-8.
3. Allen, M. J. & Yen, W.M. (1979). *Introduction to measurement theory*: Monterey, California: Brooks/Cole.
4. Al-Ananbeh, Enaam, Fathieh Abu-Moghli, and Inaam Khalaf. "PERCEPTION AND EXPERIENCES OF NURSES REGARDING PATIENTS INVOLVEMENT IN HEALTH CARE DECISIONS: AN INTEGRATIVE REVIEW." *International Journal of Applied and Natural Sciences (IJANS)* 7 (2018): 99-110.
5. Atulgan, H., & Tezbaşaran, A. A. (2005). *An investigation on consistency G and Phi coefficients obtained by generalizability theory alternative decision study for scenarios and actual cases*. *Eurasian Journal of Educational Research*, 18, 28-40.

6. Atilla, Y. (2012). Dependability of job performance ratings according to generalizability theory. *Journal of Education and Science*, 163(37), 157-164.
7. Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgment in a performance test of foreign speaking. *Language Testing*, 12, 238-257.
8. Baird, J. A., Greateorex, J., & Bell, J. P. (2004). What makes marking reliable? Experiments with UK Examinations. *Assessment in Education*, 11(3), 331-348.
9. Bergee, M. J. (2007). Performer, rater, occasion and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education*, 55(4), 344-358.
10. Breland, H., Bridgeman, B., & Fowles, M. E. (1999). Writing assessment in admission to higher education: Review and framework (ETS RR-99-3). Princeton, NJ: ETS.
11. Brennan, R. L. (2001). An essay on the history and future reliability from the perspective of replications. *Journal of Educational Measurement*, 38, 295-317.
12. Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys listening and writing tests. *Educational and Psychological Measurement*, 55, 157-176.
13. Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, H. (1972). *The dependability of behavioral measurements*. New York: Wiley.
14. Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). TOEFL[®] 2000 writing framework: A working paper (TOEFL Monograph No. MS-18; ETS RM-00-05). Princeton, NJ: ETS.
15. DHAR, SUPARNA, and MOHAMMED NAVED KHAN. "Developing a Scale to Measure Social Capital Generated and Business Benefits on Social Networking Sites." *International Journal of Human Resource Management and Research* 8.1 (2018): 1-12.
16. Feldt, L. & Brennan, R. (1989) 'Reliability', *Educational Measurement* (3rd Edition, R Linn Ed), 105-146: New York: MacMillan.
17. Fulcher, G. (2003). *Testing second language speaking*. Essex, England: Pearson Professional Education.
18. Gao, X. & Brennan, R.L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement In Education*, 14, 191-203.
19. 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(4), 323-342.
20. Gautam, Divya, and Parul Jhajharia. "The Effect of Workplace Spirituality on Employee's Self-Empowerment." *International Journal of Human Resource Management and Research (IJHRMR)* 6.3 (2016): 13-20.
21. Gronlund, N. E. (1971). *Measurement and evaluation in teaching* (3rd ed.). New York: Macmillan publishing Co.
22. Heather, C. H., Charalambos, Y. C. & Mathew, A. K. (2012). When rater reliability is not enough: Teacher observation system and a case for generalizability study. *Educational Research*, 41, 56-65. Retrieved on 12/6/2013 from <http://er.aera.net>
23. Henning, G. (1996). Accounting for nonsystematic error in performance testing. *Language Testing*, 13, 53-61.
24. Kan, A. (2007). Effect of using a scoring guide on essay scores: Generalizability theory. *Perceptual and Motor Skills*, 105, 891-905.

25. Kim, S. C & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generalizability theory. *Journal of Applied Measurement*, 10(4) 408-423.
26. Lee, Y. (2005). "Dependability of scores for a new ESL speaking test: "Evaluating prototype tasks". ETS TOEFL Monography series MS-28; USA: Educational Testing Service.
27. Lin, C. (2014). Issues and challenges in current generalizability theory applications in rated measurement. Retrieved on 16/01/15 from <https://www.ideals.ilinois.edu/.../chih-ka...>
28. Miller, M. D., & Linn, R. L. (2000). Validation of performance based assessments. *Applied psychological measurement*, 24, 367-378.
29. O'Brien, R. M. (1990). Estimating the reliability of aggregate-level variables based on individual-level characteristics. *Sociological Methods and Research*, 18, 473-504.
30. Onunkwo, G. I. N. (2002). *Fundamentals of educational measurement and evaluation*. Owerri: Cape Publishers Int'l Ltd.
31. Okeke, CHINEDU IO, CHARITY C. Okeke, and CHRISTIAN S. Ugwuanyi. "Intervention strategies that can support young adults' transition into positive fatherhood: implications for Science, Technology, Engineering and Mathematics Education." *International Journal of Mechanical and Production Engineering Research and Development (IJMPERD)* 10.3 (2020): 8585-8596.
32. Reiner, C. M., Bothell, T. W., Sudweeks, R. R., & Wood, B. (2002). *Preparing effective essay questions: A self-directed workbook for educators*. New York: New Forums Press.
33. Shavelson, R. J., & Webb, N. M. (2005). *Generalizability Theory*. Retrieved on 23/03/2010 from <http://www.web.standard.edu/.../>
34. Shavelson, R. J., Baxter, G.P., & Gao, X. (1993). Sampling Variability of Performance Assessments. *Journal of Educational Measurement*, 3(30), 215-232. Retrieved on 14/6/2014 from <http://links.jstor.org/sici?sici=0022-0655%28199323%2930%3A3%3C215%3ASVOPA%3E2.0.CO%3B2-M>
35. Tavakol, M. & Brennan, R. L. (2013). Medical education assessment: a brief overview of concepts in generalizability theory. *International Journal of Medical Education*, 4, 221-222.
36. Yang, Hui-Ling, and Wei-Pang Wu. "The Effects of Consumers' Belief regarding Internet Rumors on Purchase Intention from Different Spreading Channels." *International Journal of Information Systems, Management Research & Development (IJISMRD)* 6.1 (2016): 1-8.
37. Yelboğa, A. & Tavşancıl, E. (2010). The examination of reliability according to classical test and generalizability on a job performance scale. *Theory and Practice in Educational Sciences*, 10(3), 1825-1854.