STUDENTS' EVALUATION OF THE QUALITY OF TEACHING USING GENERALISABILITY THEORY: A CASE OF A SELECTED UNIVERSITY IN GHANA

F. Quansah
Department of Education and Psychology
University of Cape Coast, Ghana
e-mail: frank.quansah1@stu.ucc.edu.gh / https://orcid.org/0000-0002-4580-0939

ABSTRACT

Students' evaluation of lecturers' quality of teaching has been a common practice in universities in Ghana and beyond. Data gathered from students are used to make vital decisions about lecturers such as promotion, training and development needs, among others. In recent times, the accuracy of students' ratings of teaching quality has been questioned by stakeholders due to several reasons. Previous studies have attempted to investigate this issue using classical test theory (CTT) which comes with its own flaws. Little attention has been paid to the applicability of Generalisability theory (GT) to students' evaluation of teaching in Africa. This study aims to assess the reliability of students' rating of teaching through the lens of GT. A three-facet partially nested random balanced design [(*r* x *i* x *o*): *l*] was adopted for this study. Student (rater), item, and occasion served as the facets and lecturer served as the object of measurement. Both G (generalisability) and D (decision) studies were conducted. The institution's evaluation questionnaire was adapted, validated and used for data collection. The sources of measurement error were accounted for by raters and lecturer-by-occasion interaction. Generally, the dependability index for the students' evaluation of teaching quality was low, signalling little trust for such data. It was recommended that a minimum of 25 students should be permitted to rate lecturers for each class using at least 20 evaluation items. Suggestions for further studies were made based on the findings.

Keywords: students' evaluation, teaching, generalisability theory, decision study, higher education, measurement error.

INTRODUCTION

The Sustainable Development Goal 4 of the United Nations highlights the need for ensuring equitable and inclusive quality education and promoting lifetime learning prospects for all categories of people. With this, most higher education institutions (HEIs) in Ghana, if not all, have made efforts in training students for the world of work (Quansah, Appiah and Ankoma-Sey 2019). This current university training appears not to be sufficient in terms of skill development (Quansah, Ankoma-Sey and Asamoah 2019), and this has mounted pressure on

HEIs to achieving quality education, especially in the areas of teaching (also see Salem 2014). The quality of teaching is a key indicator in assessing the performance of any educational institution. It is obvious that, as key stakeholders (such as parents, students, and industry players) demand for high-quality teaching, on one hand, the management of universities also desires for an excellent measure of quality instruction (Feistauer and Richter 2016). The concept of teaching effectiveness is multifaceted and includes how instructors organise course content and communicate to learners (Staufenbiel, Seppelfricke and Rickers 2016). In a few instances, however, evidence from statistical procedures has come up against the multidimensional nature of teaching effectiveness (Li et al. 2018); this happens when the scale fails the discriminant validity test. It is worth noting that whichever way is appropriate, only if adequate evidence exists for the decision made.

Students evaluating the quality of teaching in higher education is a common phenomenon in universities around the globe (see Spooren 2010; Rantanen 2013). In most cases, students' evaluation questionnaires are developed by the institution and administered to learners after a period of instruction to assess the quality of teaching and/or learning experiences (Marsh et al. 2009; Ginns, Prosser and Barrie 2007; Staufenbiel 2000). These data from students are used to make vital decisions which can be formative and/or summative in nature (Gravestock and Gregor-Greenleaf 2008). Data used for formative purposes inform professional training needs, instruction, and the extent of students' learning. In summative decisions, data gleaned are used to inform promotion/demotion, accountability and salary structure (Iyamu and Aduwa-Oglebaen 2005; Spooren, Brock and Mortelmans 2013). The university selected for this study uses data from students' evaluation of teaching for summative and formative purposes.

From the preceding paragraph, it is clear that the importance of evaluation data, for administrators of HEIs, cannot be underestimated. Such data are used to make vital decisions and thus, the quality of the data should not be compromised (Hativa 2013). That is to say, that bad decision will be made by administrators of HEIs if students provide an inaccurate evaluation. Assuming if a group of students rate an effective teacher poorly, administrators of such institution might demote or offer development training/workshop for such an instructor which may lead to frustration and/or a waste of resources. The reverse can also result in producing academically weak students because an ineffective teacher will be blindly maintained and promoted.

Owing to the high stakes attached to the use of evaluation data from students, there appears to be tension between students (i.e. raters) and instructors (rated) (Machingambi and Wadesango 2011). On several instances, university lecturers have reiterated the need for higher education administrators to utilise students' evaluation data for only formative purposes. These

lecturers have strongly advocated against the utilisation of evaluation data for summative purposes (Manary et al. 2013; Hativa 2013). Although insufficient data are provided by the lecturers to support their position, their concerns are seen in the following questions: "Are learners competent enough to evaluate lecturers? Do the ratings of students really mirror instructional effectiveness?"

Other schools of thought have also argued against the position of those who have little trust for students' evaluation and have indicated that students are capable of accurately evaluating the quality of their learning experiences, courses and/or instruction (Hativa 2013; Benton and Ryalls 2016; Oermann et al. 2018). They further argue that contextual indicators, like poorly designed items, unfavourable occasion and students' perception of lecturers, which might result in inconsistent ratings and thus, when these factors are addressed, students will provide quality data. These scholars believe that if students are unable to provide accurate results, then, administrators in charge of quality assurance are not up to their task.

The dependability of students' evaluation of courses and/or lecturers teaching has been extensively investigated. There appear to be several inconsistencies in the results of previous studies. Whereas some studies found out low validity of students' evaluation data (e.g. Feistauer and Richter 2016; Goos and Salomons 2016; Li et al. 2018), others reveal that students' evaluation data can be highly valid (see VanLeeuwen, Dormody and Seevers 1999; Samian and Noor 2012; Ogbonnaya 2019). One major reason accounting for the mixed results centres on the approach used by these investigators. A large majority of previous studies employed Classical Test Theory (CTT) approach (see Samian and Noor 2012; Shin and Raudenbush 2012; Casabianca, Lockwood and McCaffrey 2015; Goos and Salomons 2016; Li et al., 2018). This approach, however, has flaws and thus, its use produces inadequate information for judgement about the reliability of evaluation data. For instance, measurement error in CTT is single and cannot be disentangled to capture specific measurement errors. To further separate the measurement errors, distinct analysis is required which also produces different reliability estimates (VanLeeuwen et al. 1999; Li et al. 2018). Assuming that students assess their respective lecturers on two occasions, CTT can separately estimate the reliability of this data one at a time based on which source of error is of interest. For consistency among raters, an inter-rater reliability estimate will be computed; for the reliability of the items, internal consistency estimate will be conducted; and for the stability of ratings across the two occasions, a test-retest reliability estimate will be employed. Each of these reliability procedures has its reliability coefficient and its analogous error. This makes it difficult to give a general assessment of the evaluation data. Unlike CTT, G-theory has a mechanism of combining all these three reliability procedures into a single estimation procedure to produce only one

reliability coefficient with its corresponding error.

The weakness of the CTT propelled other scholars (e.g. VanLeeuwen et al. 1999; Feistauer and Richter 2016; Li et al. 2018) to adopt generalisability theory (GT) approach, in their investigation on the validity of students' evaluation data, to liberalise the single undifferentiated error in CTT. In their study in New Mexico, VanLeeuwen et al. (1999) had item, class and student as facets and found the largest error variance to be the residual (*i* x *c* x *s*). For Feistauer and Richter's (2016) study in Germany, students, courses and teachers were the facets and teacher-by-student interaction (*t* x *s*) was revealed to have the largest variance. Li et al. (2018), in China, further employed time, student, course major type and curriculum as facets and discovered that occasion had the highest variance proportion. Li et al. (2018), and Feistauer and Richter (2016) failed to incorporate item facet although items played an important role in their study. That is, the evaluation was done using several items that the students responded to, indicating the extent to which the trait (teaching ability) was present. Additionally, VanLeeuwen et al. (1999) failed to clarify whether the class facet, which served as the object of measurement, was an evaluation of lecturers or courses. This gives unclear information on the interpretations of the results. These flaws in the previous studies which employed GT necessitates for a study of this nature and the first of its kind in Africa using this particular approach.

This current study, therefore, employs student (rater), occasion, and item as facets. Occasion was used as a facet because the investigator had observed that students' perception of lecturers seem to change after quiz results were released. This is supported by other scholars who found out that occasion was a source of measurement error (Wolfer and Johnson 2003; Casabianca et al. 2015). Lecturers were the objects of measurement. It must be noted that, in the selected university, students do not evaluate courses or learning experiences; they rather evaluate the quality of teaching, with emphasis on the activities of the instructor. The choice of the university was based on the fact that evaluation data from the students were used by the university to make vital decisions in the institution. The overarching aim of this study is to examine the dependability of students' evaluation of lecturers' teaching in a selected university in Ghana. The identity of the university remained anonymous for ethical reasons. The objectives of the study were to: (1) identify the sources of measurement error in students' ratings of teaching quality, (2) examine the accuracy of students' rating of lecturers' quality of teaching, (3) explore decisions which can be taken to improve the validity of students' evaluation. This present study expands on the knowledge from previous studies by building on the weaknesses identified and to demonstrate the applicability of GT to students' evaluation of teaching quality in the African context. The findings of the study provide a clear framework for administrators

of higher education on the utilisation and utility of students' evaluation data.

This study contributes to the literature in the area of quality and quality assurance in HEIs. Students are partners in ensuring quality in HEIs, and as such their involvement in decisions and assessments of the quality indicators cannot be under-emphasised. Several HEIs, if not all, uphold evaluation data provided by students regarding the quality of programmes run by these institutions as well as instructional tenacity. This study provides a well-balanced scholarly and comprehensive information on how student evaluation data should be used and understood in HEIs not only in Ghana but also in HEIs in Africa and beyond. To ensure that effective quality assurance decisions are made in HEIs, it important to understand the kind of data offered by significant stakeholders, such as students, on the quality of teaching and learning structures and systems.

Generalisability theory

Generalisability theory (GT) is a statistical theory on the accuracy of behavioural measurements. With regard to students' evaluation of teaching quality, the students observe the instructor for a while and further does the rating of how well the instructor delivered. The performance of the instructor is behavioural and thus, the student rater is to judge the quality of the performance based on some indicators. This provides a context within which the theory applies to this current study. GT is an extension of CTT. As a matter of fact, the concept of GT cannot be understood when segregated from CTT. In the CTT, an observed score (i.e. rating score) denoted by X is a linear function of a hypothetical true score (i.e. the expected rating) labelled as T and an error score (i.e. the difference between X and T) denoted by E. This has been shown in equation 1.

$$X = T + E \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad \text{[Equation 1]}$$

GT was developed as a result of the inefficiencies of the CTT. Stated differently, the limitations of the CTT were liberalised by GT. Most especially, GT disentangles the single undifferentiated error in CTT by the introduction of Analysis of Variance (ANOVA) procedures (see Lord and Novick 1968; Briggs and Wilson 2007) (see equation 2). In the separation of the error variance constituents of CTT, GT employs the experimental design model.

$$X = T + E_1 + E_2 + E_3 \ldots\ldots\ldots + E_n \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad \text{[Equation 2]}$$

The vocabulary of GT highlights key elements in its approach. First, the object of measurement

is the feature of the evaluation object. In the case of this study, the lecturers were the objects of measurement since their performances are rated. Any potential source of measurement error is referred to as a facet. Earlier discussions alluded to the fact that the time in which the data was obtained can produce an error to the ratings. This is reflected in the fact that students' rating behaviour is likely to change after they receive their assessment (quiz) scores. Thus, further rating scores can be influenced by their performance (i.e., whether they performed well or not).

Generally speaking, the expected rating score of the raters is dissimilar from the observed rated score. This happens because several factors come into play to account for the difference. It must be noted that the expected ratings are derived based on all possible conditions of the facet(s). The observed score, however, is derived from the sampled facet conditions. The discrepancy between the observed rating score and the expected rating score is computed by some statistical means to produce the measurement errors (relative error, δ_{pI}; absolute error, Δ_{pI}) (Shavelson and Webb 1991).

The GT framework comprises two fragments of study: Generalisability (G) study and Decision (D) study. The intent of the G study is to estimate the variance components based on various sources of measurement error. The D study, on the other hand, purports to estimate the error variances, universe score variance, and reliability coefficients using data from the G study. In essence, D study can only be conducted after the G study has been carried out first (Brennan 2001).

The use of GT helps investigators in building a context between evaluation objects and the facets of measurement depending on diverse conditions and factors admissible to the investigator (Hill, Charalambous and Kraft 2012). A high-reliability estimate (be it dependability index or generalisability index) shows that the obtained scores from the measurement can be generalised over the specified facets (Spooren, Mortelmans, and Christiaens 2014). The generalisability index ($E\rho^2$) is similar to the correlation coefficient in CTT and is norm-referenced in nature. The dependability coefficient (ϕ) is interpreted in the light of a criterion-reference test. For this study, the emphasis is placed on the dependability index.

As have been earlier underscored, GT is in tandem in its applicability to students' evaluation of courses and/or teaching. Thus, GT can generate results on whether data obtained from students are a valid measure of the actual performance of the instructors' teaching ability. Also, GT allows for the identification of several errors of measurement in the students' ratings of teaching quality to be studied. In this present study, the utilisation of GT in the analysis of the quality of students' rating of teaching was demonstrated using a sample of students from a selected university in Ghana.

METHODS

Approach

A three-facet partially nested random balanced design was adopted for this study. The design is three facets because *students* rated teaching on all the *items* on *two occasions*. Therefore, the student (rater), item, and occasion served as the facets. The student, item and occasion facets were denoted by *r*, *i*, and *o* respectively (i.e. *r* x *i* x *o*). Lecturers served as the objects of measurement and were denoted by *L*. These three facets were nested in each class (lecturer) indicating that students rated the quality of teaching on two occasions based on the items provided (i.e. (*r* x *i* x *o*): *l*). The design was random since the conditions of measurement in all the facets were random samples from the entire universe of observation. The equal number of raters from each class explains the balanced nature of the design. In the model, all negative variances were treated as zero.

Participants

The study was conducted using students from a selected university in Ghana. Thirty (30) lecturers teaching similar courses in the education department were purposively selected for the study. In other words, thirty classes were selected. None of the selected lecturers taught more than one course for the classes selected in that particular semester. Twenty students were selected from each class using a systematic sampling technique, which involves a selection procedure using an ordered sampling frame. This technique was applied by obtaining the class list with names alphabetically arranged. An interval for the selection was calculated in each class depending on the class size. A random approach was used to start the selection and subsequent selections were based on the intervals. At the end of the survey, 600 students participated in the study. The male participants were 65.7 per cent (n=394) whereas 34.3 per cent constituted the female participants (n=206).

Measures

An existing teaching evaluation questionnaire used by the selected university was adapted and used for this study. Items which were on whether the lecturer was punctual, regular, and came to class on time were deleted. This is because the emphasis of the evaluation was on the quality of teaching and thus, punctuality, for example, was not conceptualised as a quality teaching indicator, at least in this study. The questionnaire was pilot-tested and validated before the main data collection. The instrument was validated using confirmatory factor analysis (Structural Equation Modelling). The validation was done using an initial sample of 200 students taken

from a department in the university. After the confirmatory factor analysis, 16 out of 24 items had factor loadings of .5 and above and were, thus, accepted to be used for the study (i.e. the items showed an adequate level of construct validity). The result from the Multitrait-multimethod analysis did not show any evidence of multidimensionality of the scale (Geiser 2012). The reliability of the scale was estimated using omega ω. A coefficient of .84 was achieved. The model fit indices were also found to be acceptable.

Data collection plan

Formal permission and approval were sought from appropriate authorities (university management, ethical review board and lecturers) before the data collection commenced. Prior information was given to the selected lecturers and their consent was sought since they were the evaluation objects. The data were taken from the various lecturer halls; the same way the university was obtaining evaluation data from students (as at the time of data collection). With the list of students for each class (sampling frame), a systematic sampling procedure was employed to select 20 students from each class. As earlier indicated the data were obtained from two occasions (i.e. immediately before the first quizzes and just before end-of-semester examinations). During the data collection, ethical issues such as informed consent, volition, confidentiality and protection of vulnerable participants were upheld.

RESULTS

A univariate generalisability modelling was used to analyse the data through the EduG software. The analysis covered both G studies as well as D studies. The variances, sum of squares and mean squares for each source of variance were computed. Dependability coefficient which shows the level of accuracy of the students' ratings of lecturers' teaching was reported. Further optimisations were done to make decisions on the best approach to obtain valid and precise data from students.

Sources of measurement errors and accuracy of students' ratings of lecturers

The study identified 10 sources of variance in students' ratings with the three major facets and the object of measurement. The object of measurement (*l*) is not considered as a source of measurement error, however.

The results, shown in Table 1, showed that the residual (*rio:l*) with 83 per cent contributed the largest error variance to students' ratings of lecturers teaching. The next highest variance was raters (students) nested in classes (*r:l*) with 4.5 per cent variance component. Due to the nested nature of the data, however, differentiation was done using the lecturer (*l*) facet with a measurement design of *l*/*rio*. The differentiation was conducted to exempt the object of

Table 1: Sources of variances in students' ratings (Analysis of Variance)

Source	SS	df	MS	Components				
				Random	Mixed	Corrected	%	SE
l	1323.930	29	45.653	0.069	0.069	0.069	11.6	0.018
r:l	723.436	570	1.269	0.027	0.027	0.027	4.5	0.003
l	9.353	15	0.624	0.000	0.000	0.000	0.0	0.000
o	1.095	1	1.095	0.000	0.000	0.000	0.0	0.000
li	215.614	435	0.496	0.000	0.000	0.000	0.0	0.001
lo	28.032	29	0.967	0.001	0.001	0.001	0.2	0.001
ri:l	3662.189	8550	0.428	-0.032	-0.032	-0.032	0.0	0.005
ro:l	269.217	570	0.472	-0.001	-0.001	-0.001	0.0	0.002
io	5.522	15	0.368	0.000	0.000	0.000	0.0	0.000
lio	241.226	435	0.555	0.003	0.003	0.003	0.5	0.002
rio:l	4201.408	8550	0.491	0.491	0.491	0.491	83.0	0.008
Total	10681.022	19199					100%	

SE – Standard Error; SS- Sum of Squares; MS- Mean Squares

measurement to find out whether there will be a change in the results. Table 2 shows the result on the differentiation.

Table 2: Differentiating L (lecturers) G Study Table (measurement design L/RIO)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% Relative	Absolute error variance	% Absolute
l	0.069		.....		.....	
	.....	***r:l***	0.001	46.5	0.001	46.0
	.....	***i***	.....		0.000	0.6
	.....	***o***	.....		0.000	0.6
	.....	***li***	0.000	0.2	0.000	0.2
	.....	***lo***	0.001	23.3	0.001	23.0
	.....	***ri:l****	(0.000)	0.0	(0.000)	0.0
	.....	***ro:l****	(0.000)	0.0	(0.000)	0.0
	.....	***io****	.....		(0.000)	0.0
	.....	***lio***	0.000	3.4	0.000	3.4
	.....	***rio:l***	0.001	26.6	0.001	26.3
Sum of variances	0.069		0.003	100	0.003	100

Generalisability index ($E\rho^2$) = 0.46 Dependability index (ϕ) = 0.46

*The sources of measurement error with negative variances

The results revealed that the rater facet (46%) contributed the largest variances in ratings (Table 2). The variance component for the residual (interaction of all sources and other systematic or unsystematic factors) was also large (26.6%). The total variability due to lecturer-by-rater interaction was relatively large (23%).

The source of variability for lecturer-by-occasion interaction was further probed to find

out the spread of variances for the two occasions (see Figure 1). It was revealed that the variances for each class on the first occasion (before the first quiz) were significantly smaller than the variances for the second occasion (before semester examination).

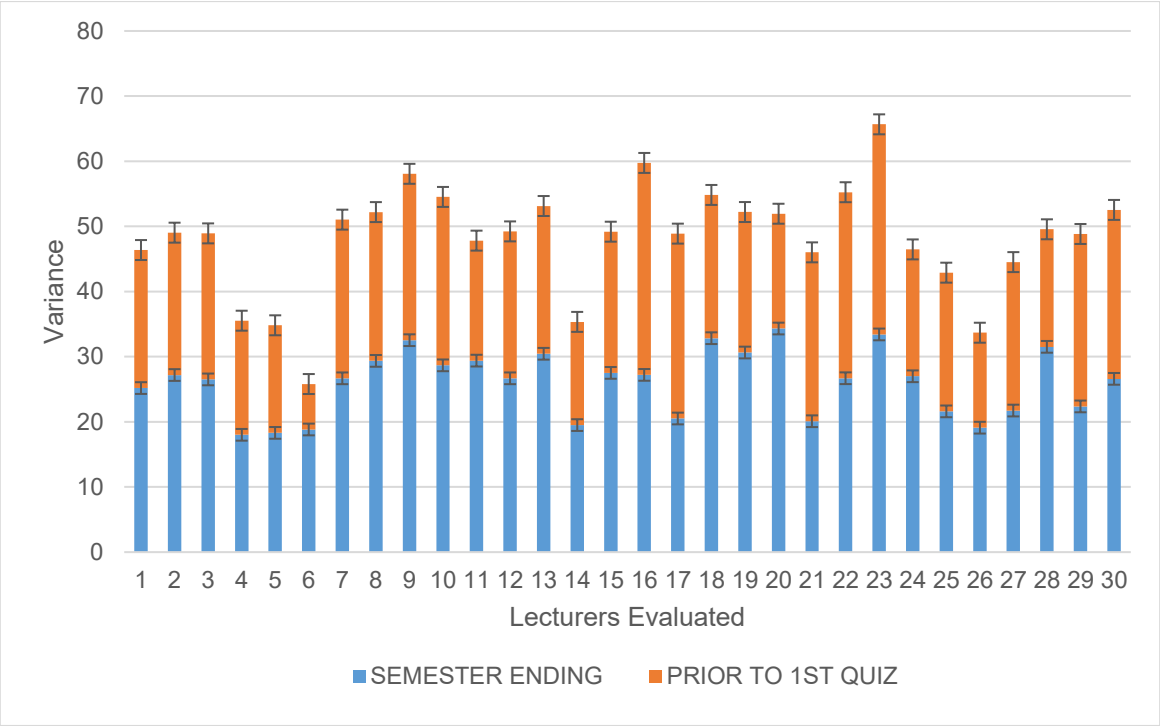

Figure 1: Variances in ratings for Time 1 (Semester ending) and Time 2 (Prior to 1st quiz)

Decision study: Decisions which can be taken to improve the quality of students' evaluation of teaching

A D (decision) study was conducted using the result from the G study. Optimisation was carried out for decisions which can be taken to improve the quality of data taken from students. The details of the result are presented in Table 3.

Table 3: Optimisation (*I*/*rio*) from D Study

	Option 1	Option 2	Option 3	Option 4	Option 5	Option 6
Observations	9000	19200	30000	45000	63000	84000
L	30	30	30	30	30	30
R:L	15	20	25	30	35	40
I	10	16	20	25	30	35
	Two occasions					
Coef_G rel. ($E\rho^2$)	0.74	0.76	0.77	0.77	0.78	0.78
Coef_G abs. (ϕ)	0.74	0.76	0.77	0.77	0.78	0.78
	One occasion					
Coef_G rel ($E\rho^2$)	0.71	0.76	0.88	0.96	0.96	0.97
Coef_G abs. (ϕ)	0.71	0.76	0.88	0.96	0.96	0.97

The results, as shown in Table 3, suggest that 15 student raters in each class using 10 evaluation items on two different occasions yields a relatively poor rating from students ($\phi = .74$). Further results discovered that 40 raters in each class rating a lecturer using 35 items on a single occasion produced more valid evaluation data. However, the difference between the coefficients of option 3 and 6 was not too large. As a result, it can be said that 25 students evaluating a lecturer using 20 items on a single occasion appear to produce a valid result just like option 6. A pictorial view of the result is shown in Figure 2.

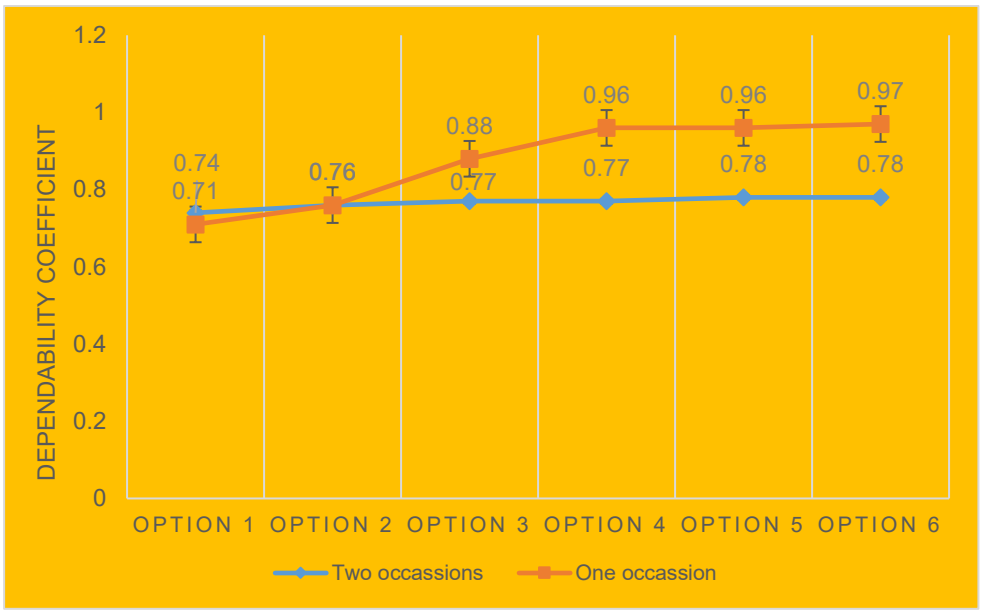

Figure 2: Trend of optimisation results (D study)

DISCUSSION

The study revealed that the relative rating score of lecturers systematically differed from one lecturer to another. This is something expected since the teaching ability or proficiency of the lecturers are likely to differ. However, the lecturer (object of measurement) is not considered as a source of measurement error. From a broader perspective, rater nested in lecturer had the highest contribution to the total variability of students' rating. In other words, student ratings for the same lecturer systematically differed from one rater to the other. This suggests that the level of students' agreement of the extent of teaching proficiency of the same lecturer was low. Students who were taught by the same lecturers had a different opinion of the quality of teaching of such lecturer. This is consistent with some previous studies who found rater variability as a source of measurement error (e.g. Feistauer and Richter 2016). This can also be attributed to the fact that students may not be clear in their might which behavioural traits constitute effective teaching and which ones reflect poor teaching. Therefore, they may be left alone to do this evaluation subjectively, although little evidence exists to this effect.

Another source of measurement error in students' ratings of lecturers' teaching quality was lecturer-by-occasion interaction. That is, the relative ratings of lecturers differed from one occasion to another. This is to say that the ratings of each lecturer systematically varied on the two different occasions. So, a lecturer on one occasion may be rated highly but rated poorly on another occasion. The implication is that it is either the raters who were largely inconsistent or the teaching proficiency of the lecturers were not stable because of factors like the topic being handled. The raters may also differ in their ratings on different occasions due to factors such as low or high scores in assessment, changes in perceptions, the nature of course been taught (Li et al. 2018). In the case of this study, ratings were obtained in two instances: prior to writing the quizzes and after being assessed. A major contributory error here could be that the assessment was difficult or too easy, or even the students' obtained high or low scores in their assessment. In instances where students received less difficult assessment tasks or high assessment scores, they are more likely to rate lecturers high and vice versa (Wolfer and Johnson 2003; Casabianca et al. 2015). There were other systematic and random sources of measurement error which this study did not identify but provided evidence that they existed.

The dependability index for students' evaluation of lecturers was low, signalling little trust for such data. It must be stated that variability in raters and lecturer-by-occasion interaction accounted for approximately 69 per cent to the students' ratings of teaching quality. This questions the use of such data for decision-making purposes, especially for summative decisions. This finding supports the observations of previous studies who found the existence of low reliability of students' ratings of lecturers' teaching (e.g. Feistauer and Richter 2016; Goos and Salomons 2016; Li et al. 2018). To improve on the dependability of students' ratings the D study suggested that data should be taken from a minimum of 25 students on one occasion with at least 20 evaluation items; this is found to produce much dependable evaluation data. However, this suggestion should be implemented bearing in mind the resources available, time for the administration, and other considerations such as the class size.

CONCLUSIONS AND RECOMMENDATIONS

Students' evaluation of teaching are not carried out for formality sake but the data are used to make vital decisions. In this study, it was found out that such data had low reliability as there was little consistency in the ratings of students regarding teaching effectiveness. The present study emphasised the significant role of variability in raters and lecturer-by-occasion interaction in explaining students' ratings of teaching. The findings from my study do not suggest that students' evaluation should not be trusted and utilised but rather should be used with caution otherwise it can compromise, motivate or demotivate lecturers and in turn promote or stifle the

quality of teaching and learning. Therefore, these results are a signpost for administrators of HEIs to pay special attention to the quality indicators of the information provided by students as this information has an overall implication on quality. Thus, this high unreliable students' rating tells the extent to which decisions can be made with the data.

The findings of this study should be interpreted with caution since the study was conducted in only one department in a single university. Hence, the applicability of the results to other departments in that particular university and other universities might be problematic. I recommend that further studies should include evaluators from diverse programmes or courses to understand the scope of the ratings. Again, an evaluation data on a particular lecturer from only a sample of students may reduce the representativeness of the ratings to the class population. It is suggested that future studies should involve all students in each class to engage in the evaluation process. This is because GT is sample dependent and the results might change from one sample to another. It was also realised that some of the lecturers evaluated had already taught the students in previous semesters prior to the semester which the study was conducted. This means that some of the lecturers had taught the students two courses in reality. Consequently, the previous performance of those lecturers might confound the students' ratings on the performance of the current course.

The following recommendations were suggested to the management of HEIs, specifically, the Quality Assurance Unit or Department:

1. There should be students' sensitisation and awareness of the need to provide an accurate evaluation of their lecturers' quality of teaching. Again, there should be a clear framework on the lecturers' expectations of teaching to inform the students' judgements of the quality indicators of teaching. These practices are to reduce the variability of students' rating.
2. An adequate number of students should be involved in the evaluation of each lecturer. A minimum of 25 students should be allowed to evaluate each lecturer.
3. Evaluation items should be sufficient to estimate the construct of teaching effectively. At least 20 items should be used for the evaluation of lecturers' teaching.

REFERENCES

Benton, S. L. and K. R. Ryalls. 2016. Challenging misconceptions about student ratings of instruction. *IDEA Paper* 58(1): 1–22.

Brennan, R. L. 2001. *Generalizability theory*. New York, NY: Springer-Verlag.

Briggs, D. C. and M. Wilson. 2007. Generalizability in item response modelling. *Journal of Educational Measurement* 44(2): 131–155.

Casabianca, J. M., J. R. Lockwood and D. F. McCaffrey. 2015. Trends in classroom observation scores.

Educational and Psychological Measurement 75(2): 311–337.

Feistauer, D. and T. Richter. 2016. How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education* 42(8): 1263–1279.

Geiser, C. 2012. *Data analysis with Mplus*. New York, NY: Guilford Press.

Ginns, P., M. Prosser and S. Barrie. 2007. Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. *Studies in Higher Education* 32(5): 603–615.

Goos, M. and A. Salomons. 2016. Measuring teaching quality in higher education: Assessing selection bias in course evaluations. *Research in Higher Education* 58(4): 341–364.

Gravestock, P. and E. Gregor-Greenleaf. 2008. *Student course evaluations: Research, models, and trends*. Toronto, Canada: Higher Education Quality Council of Ontario.

Hativa, N. 2013. *Student ratings of instruction: A practical approach to designing, operating, and reporting*. Tel Aviv: Oron Publications.

Hill, H. C., C. Y. Charalambous and M. A. Kraft. 2012. When rater reliability is not enough: Teaching observation systems and a case for the generalizability study. *Educational Researcher* 41(2): 56–64.

Iyamu, E. O. S. and S. E. Aduwa-Oglebaen. 2005. Lecturers' perception of student evaluation in Nigerian Universities. *International Education Journal* 6(5): 619–625.

Li, G., G. Hou, X. Wang, D. Yang, H. Jian and W. Wang. 2018. A multivariate generalizability theory approach to college students' evaluation of teaching. *Frontiers in Psychology* 9(1065): 1–11.

Lord, F. M. and M. R. Novick. 1968. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Machingambi, S. and N. Wadesango. 2011. University lecturers' perceptions of students' evaluation of their instructional practices. *Anthropologist* 13(3): 167–174

Manary, M. P., W. Boulding, R. Staelin and S. W. Glickman. 2013. The patient experience and health outcomes. *The New England Journal of Medicine* 368(3): 201–203.

Marsh, H. W., B. Muthén, T. Asparouhov, O. Lüdtke, A. Robitzsch, Alexandre J. S. Morin and U. Trautwein. 2009. Exploratory structural equation modelling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modelling: A Multidisciplinary Journal* 16(3): 439–476.

Oermann, M. H., J. L. Conklin, S. Rushton and M. A. Bush. 2018. Student evaluations of teaching (set): Guidelines for their use. *Nursing Forum* 53(3): 280–285.

Ogbonnaya, U. I. 2019. The reliability of students' evaluation of teaching at secondary school level. *Problems of Education in the 21st Century* 77(1): 97–109.

Quansah, F., V. R. Ankoma-Sey and D. Asamoah. 2019. The gap between the academia and industry: Perspectives of university graduates in Ghana. *International Journal of Education and Research* 7(3): 63–72.

Quansah, F., E. Appiah and V. R. Ankoma-Sey. 2019. University students' preparation towards building knowledge economy in Africa: A case of universities in Ghana. *International Journal of Social Sciences & Educational Studies* 6(1): 38–48.

Rantanen, P. 2013. The number of feedbacks needed for reliable evaluation: A multilevel analysis of the reliability, stability and generalisability of students' evaluation of teaching. *Assessment & Evaluation in Higher Education* 38(2): 224–239.

Salem, M. I. 2014. The role of universities in building a knowledge-based economy in Saudi Arabia. *International Business & Economics Research Journal* 13(5): 1047–1056.

Samian, Y. and N. M. Noor. 2012. Students' perception on good lecturer based on lecturer performance assessment. *Procedia-Social and Behavioural Sciences* 56(1): 783–790.

Shavelson, R. J. and N. M. Webb. 1991. *Generalizability theory: A primer*. Thousand Oaks, CA: Sage Publications.

Shin, Y. and S. W. Raudenbush. 2012. Confidence bounds and power for the reliability of observational measures on the quality of a social setting. *Psychometrika* 77(3): 543–560.

Spooren, P. 2010. On the credibility of the judge: A cross-classified multilevel analysis on students' evaluation of teaching. *Studies in Educational Evaluation* 36(4): 121–131.

Spooren, P., B. Brockx and D. Mortelmans, 2013. On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research* 83(4): 598–642.

Spooren, P., D. Mortelmans and W. Christiaens. 2014. Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Studies in Educational Evaluation* 43(1): 88–94.

Staufenbiel, T. 2000. Students course assessment questionnaire for evaluation of university courses. *Diagnostica* 46(4): 169–181.

Staufenbiel, T., T. Seppelfricke and J. Rickers. 2016. Predictors of student evaluations of teaching. *Diagnostica* 62(1): 44–59.

Wolfer T. A. and M. M. Johnson. 2003. Re-evaluating student evaluation of teaching. *Journal of Social Work Education* 39(1): 111–121.

VanLeeuwen, D. M., T. J. Dormody and B. S. Seevers. 1999. Assessing the reliability of student evaluations of teaching (SETS) with generalizability theory. *Journal of Agricultural Education* 40(4): 1–9.