Heliyon 8 (2022) e12548

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Item and rater variabilities in students' evaluation of teaching in a university in Ghana: application of Many-Facet Rasch Model

Frank Quansah *

Department of Educational Foundations, University of Education, Winneba, Ghana

ARTICLE INFO

Keywords:
Item
Rater
Students' evaluation
Variances
Many-facet model
Teaching

ABSTRACT

This research examined the item and rater variabilities in students' evaluation of teaching and courses exercise at the University of Cape Coast (UCC) through the lenses of the Many-Facet Rasch Model (MFRM). The study covered students during the 2019/2020 academic year in the selected university, analysing secondary data obtained from the Directorate of Academic Planning and Quality Assurance, UCC (DAPQA-UCC). The data were analysed by conducting partial credit MFRM analyses. It was found that the sources of measurement errors in the student evaluation exercise included halo effect, non-functional item structure, inconsistent students' ratings, rater leniency, and non-functional rating scale. It was concluded that data from students' appraisal of lecturers' teaching should be used with caution. It was recommended that DAPQA-UCC and the university management should train students on the evaluation of teaching, as well as review the existing evaluation form for appraising courses and teaching by subjecting the instrument to rigorous validation procedures.

1. Introduction

In the affairs of higher education, students' appraisal of courses and teaching quality is common in almost every tertiary institution across the globe (Rantanen, 2013). Other means of evaluation, like teacher certification and peer assessment, have also been mentioned in the literature; however, they are uncommon in terms of the evaluation of teacher quality (Becker et al., 2011). In fact, students' appraisal of courses and teaching, regardless of other approaches, are employed on a large scale in higher education institutions to evaluate the instructional quality and to compare instructor performances across departments, courses and, sometimes, among universities (Becker et al., 2011). Arguably, students' evaluation of courses and teaching influences the promotion of instructors (Galbraith et al., 2012), students' university application (Alter and Reback, 2014), and students' choice of courses (Wilhelm, 2004). In countries like the United States, students' evaluation data are used for official and unofficial ranking of institutions, and auditing purposes (Johnson, 2000). These uses of students' evaluation data have sparked extensive scientific literature in areas such as psychology, education, economics and sociology (Goos and Salomons, 2017). With this understanding, the central issue baffling researchers is the extent to which students' evaluation data can be understood as a pointer for examining the quality of teaching in higher education (Taut and Rakoczy, 2016).

Applying the Kirkpatrick evaluation model in explaining the dynamics of students' appraisal of teaching, Kirkpatrick (1959) highlighted that students evaluate courses and teaching based on four indicators, namely, *Reaction, Learning, Behaviour, and Results.* If students enjoy the course the instructor is handling, and as well believe that much has been learnt, they are likely to provide good ratings (reaction dimension). Also, students will rate instructors excellently if they can demonstrate what they have learnt in the short term in situations such as class tests (learning dimension). At the behavioural level, the students will rate instructors with distinction if they can apply what they have learnt to real-life situations. Lastly, the result dimension reflects the point where the students benefit from the course in the long term such as securing employment due to the skills acquired from the course. The result dimension does not directly feature in students' ratings at the school level since the students would still be in school at the time of the evaluation. Despite the relevance of the Kirkpatrick model, critics have stressed that the ratings by students in evaluation exercises are mostly limited to the lower levels of the model (i.e., Reaction and Learning) and sometimes based on other unrelated aspects of instruction due to differences in students' perspectives (Steele et al., 2016).

The validity of students' evaluation of courses and teaching is a contentious issue (Hornstein, 2017). In an extensive review by Spooren et al. (2013), it was established that there are conceptual, theoretical and

* Corresponding author.
E-mail address: fquansah@uew.edu.gh.

https://doi.org/10.1016/j.heliyon.2022.e12548
Received 13 June 2022; Received in revised form 16 November 2022; Accepted 14 December 2022

empirical supports for students' appraisal of courses and teaching. Despite these pieces of evidence, the utilisation of data from students' evaluation of teaching as a proxy for the quality of teaching and course contents has also been critiqued for several reasons (Spooren et al., 2013). In particular, it has been argued that data from students' appraisal of teaching are polluted by noise. This presupposes that students have been found to evaluate instructional quality based on some characteristics of the course (e.g., the difficult/easy nature of the course), students (e.g., students being friends with the instructor or dislike for the course) and teacher (e.g., the strictness of the instructor) which are unrelated to the quality of teaching and course contents (Ko et al., 2013). Such contamination in the evaluation process results in situations where instructors have been found to inflate the grades of students; literature has found a direct link between students' evaluation of teaching and their scores/performances, regardless of the learning outcomes (Ewing, 2012).

Several other scholars have lamented over the validity of data from students' appraisal of courses and teaching, especially concerning their capacity to evaluate their teachers (e.g., MacNell et al., 2015; McPherson et al., 2009). Such laments, nevertheless, are addressed by indicating that once there is some sort of accurate information provided by the students in their evaluation of teaching quality, there can be adjustments in the data to control for extraneous variables (McPherson and Jewell, 2007; McPherson et al., 2009). Other scholars, like Ogbonnaya (2019), have supported this approach to evaluation and have adjudged it as a suitable and appropriate means of evaluating teaching effectiveness. Experts have shown that students are eligible and capable of appraising the degree to which instruction is satisfactory, informative, worthwhile or productive (McPherson et al., 2009). This view was supported by Theall and Franklin (2001) who noted that students' evaluation of how much is learnt in a course is highly related to their overall appraisal of teaching quality and effectiveness.

A review of existing literature on issues of validity and dependability of students' appraisal of courses and teaching around the globe has shown the the predominant utilisation of three major measurement approaches: Classical Measurement Theory (CMT), Generalizability Theory (GT) and Many-Facet Rasch Model (MFRM). The majority of these previous studies employed CMT which is the foundation of most, if not all, measurement theories (e.g., Adams and Umbach, 2012; Braga et al., 2011; Fah and Osman, 2011; Goos and Salomons, 2017; Ogbonnaya, 2019; Samian and Noor, 2012; Sliusarenko, 2013; Raza and Irfan, 2018; Zhang et al., 2015). It is in very few instances that studies explored the validity and reliability of students' evaluation of teaching using the GT approach (e.g., Feistauer and Richter, 2016; Li et al., 2018; Quansah, 2020; VanLeeuwen et al., 1999). A single study, however, investigated the dependability of students' appraisal of courses and teaching through the lens of MFRM (Börkan, 2017).

Research has shown that the MFRM approach to investigation within the framework of rater-mediated assessment provides better and more comprehensive information relative to the CMT and GT approaches (Linacre, 1989). Comparing the CMT to the MFRM, the CMT analyses do not provide explicit information regarding the reliability of the ability levels of the objects of measurement (i.e., lecturers in this study). On the contrary, the MFRM analysis reports detail information about the estimation of the ability levels of the objects of measurement (Goffin and Olson, 2011). Whereas analyses in CMT directly utilise the raw scores from the dataset, facet/factor measurements are converted to a wide metric scale, known as logit, in the MFRM (Linacre, 1989). Similarly, studies conducted using the GT approach, as compared to the MFRM, also do not provide a detailed picture of the accuracy of students' appraisal of courses and teaching (Börkan, 2017; Linacre, 2003). GT offers an overall summary of group-specific analysis of the variances influencing the scores for all objects of measurement (e.g., lecturers, as in this study) (Linacre, 1994). Unlike GT which has no implication for the individual object of measurement, except the number of observations made, MFRM focuses on the individual object of measurement and as well provides quality control statistics for each element on a common and equal-interval scale (Linacre, 1994). While GT, for example, can describe how different lecturers were rated by the students, it cannot provide information on the fairness dynamics of the rating (i.e. which lecturers were fairly rated and which ones were not rated fairly). Whereas MFRM can provide a fair score distribution for each object of measurement, GT cannot (Linacre, 1994). Given this background, the present study sought to examine the item and rater variabilities in students' evaluation of teaching exercise using UCC as the study setting. Specifically, the study sought to assess: (1) the behaviours of students (e.g., halo effect and severity effect) in the appraisal of courses and teaching; and (2) the effectiveness of the evaluation items in the measurement of the quality of course content and teaching.

1.1. Context of the study

In the University of Cape Coast (UCC), data from students' evaluation of the quality of courses and teaching serve both formative purposes (i.e., lecturer development decision, and improvement and/or modifications in teaching and course contents) (UCC, 2012) and summative purposes (i.e. lecturers' appointment and promotion) (UCC, 2015). While little evidence has been established to argue out that a lecturer may be denied an appointment or promotion based on students' appraisal of teaching, it is obvious that this happens, as a matter of policy (UCC, 2015). Given the significant uses of this evaluation data, its quality is paramount and this quality concern becomes an issue not only to university administrators but also to instructors and researchers.

In a cross-sectional survey among lecturers of all ranks at UCC, the lecturers lamented that students poorly rated the instructors who were strict in their dealings (Gyimah et al., 2016). Although the recognition and acceptability of students' appraisal data were high among the lecturers, some of them argued that the students do not have a good value-judgment in assessing their teaching activities and thus, the students may not be responsible enough to take up this evaluation role (Gyimah et al., 2016). In the researcher's view, claims of this nature by a cross-section of the lecturers are worrying and raise several questions about the tenacity of the data the students provide. This concern was supported by another study conducted in UCC by Kwarteng, Doku, Matta, and Doh-fia (2014) which reported fluctuations in lecturers' overall ratings provided by the students over 5 years (i.e., from the 2008/2009 academic year to 2012/2013). These studies raise concerns regarding the dependability of the data obtained from the students for decision making in the said university.

2. Materials and methods

2.1. Study design

The three-facets design within the MFRM framework was employed for the conduct of the study. The facets were lecturer (i.e., object of measurement), student (i.e., rater) and item. It must be emphasized that the object of measurement, which is the lecturer, is considered a facet in the context of MFRM. This study adopted partial credit modelling (PCM) because the instrument used in gathering the data by the university has different response categories across the evaluation items. The PCM is adopted in situations where the number of response categories across the items is different, or the relative difficulty between the response categories is likely to differ from one item to another (Masters, 2010).

2.2. Sample

The study used secondary data. These data are information already obtained from the students by the university for decision-making purposes. Through a purposive sampling technique, 145 courses were selected with no students having duplicated responses. In all 2553 students' responses for the 2019/202 academic year were obtained from the Directorate of Academic Planning and Quality Assurance (DAPQA), UCC.

2.3. Description of the evaluation form

This research relied solely on the data retrieved by the DAPQA-UCC using an already existing questionnaire developed by the university. This evaluation form used has 25 items comprising 22 closed-ended and 3 open-ended items. The closed-ended items are used for assessing five dimensions: course outline (1 item), course content (3 items), attendance (3 items), mode of delivery (10 items), and assessment (5 items). In this study, the course outline dimension was removed from the data due to three reasons: (1) DAPQA-UCC do not make use of this dimension in determining the lecturers' level of effectiveness in handling the course; (2) the item under this section is factual such that the variability of the responses is attributed to forgetfulness or mere sabotage; (3) the dimension has only one item and this presupposes there are no conditions of measurement and for that matter, that item ceases to be a facet. Also, two other sections (i.e., attendance and assessment domains) on the appraisal evaluation form were not included in this research. These two dimensions comprised items that were objective and did not require any subjectivity in the scoring resulting in approximately zero variances. With regards to the 3 open-ended items, students were required to write the lecturer's strength(s), weakness(es), and any other suggestions. The responses to these open-ended questions were also not included in this research because the responses provided were not quantitative. Only two sections of the evaluation form were involved in this study, these are, the mode of delivery and course contents domains. The items for these two sub-scales were 13 in all.

2.4. Ethical considerations and data management

The study took into consideration ethical issues such as the Ethical Review Board (ERB) clearance, confidentiality, anonymity, and data de-identification. The researcher applied for ethical clearance from the ERB of UCC. Also, confidentiality was not compromised, in the sense that the details of the data were not communicated to any third party. That is, the researcher ensured that the data did not leak to other parties who may need it (Tripathy, 2013). The data extracted had no traces of students' identification in relation to a particular set of responses. All indicators of students' personal characteristics were removed from the data as a way of anonymizing their identities. In addition, the identity of the lecturers/courses evaluated was also anonymized and replaced with pseudonyms through a de-identification process (e.g., AA, AB, AC, AD, AE, etc.)

Informed consent was provided by DAPQA-UCC. This is because students during the administration phase permits the university to use the data for managerial and administrative purposes as well as for future research (Jol and Stommel, 2016). Thus, informed consent has been already provided retrospectively and this provision empowers the university to give the data to any party for other purposes stated within institutional rules. The data were acquired in a soft copy and were validated manually. The validation process included checking for the accuracy of data and ensuring that the ratings provided fell within the rating scale. The data were stored on a laptop with a password that was only known to the investigator. The data were processed and presented in a manner that had no traces of the responses of the students.

2.5. Statistical analyses

The data for the study were analysed using the MFRM and processed with the FACET computer programming software. The MFRM is an extension of Rasch analysis and thus, operates as a multi-dimensional item response theory (Linacre, 1994). MFRM calibrates each facet using the same logit linear scale after correcting the raw scores for variations among rater severity, and variances in the relative difficulty of the task (Lunz et al., 1990). In this study, the person facet was positively arranged (i.e., high on the logit scale denotes high ability), the rater facet was negatively arranged (high on the logit scale means more severe ratings and vice versa), and the item facet was also negatively ordered (i.e., high on the logit scale represents more difficult items). The analysis of MFRM provided the following indices and assessment criteria:

1. The Infit and Outfit mean square (MnSq) estimates: The infit and outfit MnSq estimates denote the difference between observed and the expected model-driven responses and flag unexpected responses in the data set (Linacre and Wright, 2002). The value of these estimates ranges from zero to infinity. In the event of perfect correspondence, the value of these estimates becomes 1. When the estimate is more than 1, then the variance is larger than expected. For rater fit indices, high variance denotes that a rater provided ratings in an unpredictable and inconsistent manner. A value less than 1 indicates low variance in the data than that predicted by the model. With respect to the rater facet, these estimates can be deduced as too predictable behaviour of the rater. That is, the rater is either too consistent in his/her ratings or is not able to discriminate between different performance levels. Linacre and Wright (2002) recommended that the infit and outfit MnSq should have an estimated value between 0.5 and 1.5. Notwithstanding this rule of thumb, Wu and Adams (2013) argued that the infit and outfit statistics are sensitive to sample size and, as result, caution should be taken when deciding on the upper and lower control limit for these estimates. They, therefore, conducted several simulations and designed appropriate lower and upper control limits for specific samples. From their recommendations, lower and upper control limits of .90 and 1.10 respectively for infit and outfit parameters are reasonable enough to make a good judgement of data with over 700 cases. This study operated with this recommendation since the sample size was more than 700.
2. The Separation Ratio (G): The separation ratio is symbolised as G. The separation ratio indicates the extent of the dispersion of the estimates, compared to their error of measurement. This statistic ranges from 1 to infinity. For example, when $G = 3$, it suggests that the statistical spread in the measures of the elements in the facet is three times more than the imprecision in their estimations (Wright, 1996). Whereas a high G value is preferred for the person (i.e., lecturer) and item facets, a low G value is preferred for the rater facet (i.e. student).
3. The Reliability of Separation Index (R): The reliability of the separation index is symbolised by R. This index illustrates how reproducibly different the measures are. The reliability of the separation index ranges between 0 and 1. If this index is close to 1, there is a greater likelihood that the elements of the facet with large measure estimates essentially have greater measures than those with low measure estimates (Linacre, 2009). The same way the G value is interpreted, a high R-value is preferred for the person (i.e., lecturer) and item facets, whereas a low R value is preferred for the rater facet (i.e., student).
4 Fixed chi-square statistics (all-same statistics): The fixed chi-square statistic is a test conducted to examine the model fit of the data. In a specific case of MFRM, a hypothesis is tested to examine whether or not the estimations of each element of a facet have an equivalent estimate after the error of measurement has been accounted for.
5. Scale functioning quality measures: The scale functioning quality measures provide information regarding the quality of the scale and scale categories. A major indicator is by inspecting the frequencies and percentages of the scale categories. Scale categories with relatively low frequencies and percentages show a non-functional rating scale and such categories need to be revised or possibly combined with other scale categories. Also, the category average measure is another indicator that can be examined to understand the quality of scale functioning. In this case, the average logit measure of a particular scale category should either be less than scale categories of higher value or greater than scale categories with lower values. For example, if the average logit measure of a scale category of 2 is 2.43, then the average logit measure of a scale category of 1 should be less than 2.43, and greater than 2.43 for a scale category of 3. Also, the outfit parameter should be closer to 1, if the scale is functioning properly.

3. Results

3.1. Students' behaviours in the appraisal of courses and teaching

The study evaluated the rating behaviours of students during the appraisal of courses and teaching. Details of the results are shown in Tables 1, 2, and 3.

Table 1 provides the summary of fit statistics for the rating behaviours of students in evaluating the quality of teaching and course contents. Although both infit and outfit statistics are provided, much emphasis will be placed on the infit statistics since outfit statistics are sensitive to outliers. The results (in Table 1) revealed that about 34.8% of the students had an acceptable fit. This result suggests that 34.8% of the student-raters provided ratings that were consistent with the true/expected ability of the lecturers/instructors.

Notwithstanding this, a significant number of students also recorded overfit (30.5%) and misfit (34.7%) for the model. The overfit statistics suggested that about 30.5% of the students who participated in the evaluation exercise were either not able to discriminate between different performance levels of the lecturers or there is the presence of halo effects. The misfit statistics showed that there was more variance in the ratings than expected among 34.7% of the students. This indicated that the students were inconsistent and unpredictable.

Table 2 further provides a summary of the rater measurement report which presents details of the entire model.

The summary statistics included a model summary and location summary of the analysis (Table 2). A closer look at the model summary showed a significant chi-square value, χ^2 (213) = 258.1, $p < .001$. The chi-square analysis tests the hypothesis that all student-raters were equivalent in their level of severity. Based on the chi-square test, the students were not equivalent in their level of severity. The separation strata index of 2 also indicated that the students were heterogeneous in rating the quality of teaching and course contents. This was also confirmed by the moderate reliability of the separation coefficient (which was further away from zero) and a separation ratio (which was slightly greater than 1). The location summary, as shown in Table 2, also revealed that the students were generally lenient in their ratings of teaching and course appraisal (logit = -1.04, S.E = .69). This was reflected in the overall observed rating of 3.63 and the expected rating of 3.51, showing that several students provided higher scores than expected. The overall infit and outfit statistics showed some elements of misfit indicating inconsistent ratings and high variances in the ratings.

In Table 3, individual cases with their detailed parameters were studied. This comprised those raters who recorded misfit as well as those who recorded overfit.

The first 5 cases were raters whose responses produced misfit and the rest are those whose responses produced overfit statistics. The results, shown in Table 3, revealed that the raters who were misfitting consistently provided low ratings than expected. Rater 12, for instance, provided a rating score of 2.62 for the instructor instead of a true/expected score of 2.77. Similarly, rater 41 rated an instructor with a score of 2.31 instead of an expected score of 2.64. Raters 15, 6, and 94 were all found to follow a similar trend of ratings; providing a score lower than the expected score of the instructor. An observation common among the raters who were misfitting is that they were all found on the positive side of the logit scale indicating that they had high severity in their ratings (from 2.44 to 2.99).

This was not the case for those raters who were overfitting; they were more lenient in their ratings. These raters were all located on the negative side of the logit scale (from -2.01 to -3.24). Consistently, overfitting raters were found to provide high ratings than expected. Rater 128, for example, scored the instructor with a rating of 3.85 instead of a true rating score of 3.34. Likewise, rater 129 provided a rating of 4.0 for an instructor who has a true rating of 3.88.

Table 1. Fit statistics for students' rating behaviours.

Fit Statistics	Indicators	Infit	Outfit
<.90	Overfit	825(30.5)*	884(32.7)
.90–1.10	Acceptable fit	940(34.8)	893(33.0)
>1.10	Misfit/Undefit	937(34.7)	925(34.3)

* percentages in parenthesis.

Table 2. Summary statistics for rater (Student) measurement report.

Model Summary	Estimate	Location Summary	Estimate
RMSE	.65	Observed rating ± SD	3.63 ± .28
Separation ratio	1.05	Expected rating ± SD	3.51 ± .27
Reliability of separation	.52	Logit measure	-1.04
Separation strata index	2.0	Model S.E.	.69
Chi-squared	258.1	Infit	1.19
df	213	Outfit	1.36
p-value	.000*		

RMSE- Root Mean Square Error; S.E- Standard Error; SD- Standard Deviation.
* significant at p < .001.

3.2. Effectiveness of evaluation items in the measurement of the quality of course content and teaching

The research also examined the effectiveness of the evaluation items in the measurement of the quality of course content and teaching. Three different aspects of the items were assessed to find out whether the: (1) structure of the item was appropriate and clear, (2) criteria were well understood and the items accurately measured a particular criterion, and (3) rating scale is of good quality and was well understood by the raters. The unexpected responses for the model were first presented to provide a general idea about the three different aspects of the items. Table 4 presents the details of the result of the unexpected responses.

The results presented in Table 4 revealed that items 8, 6, 5, 1, 9, and 3 had problems. These items were further investigated to identify the problem at hand. Also, the criteria labelled mode of delivery (teaching quality) was flagged as problematic and thus, there was the need for further investigation (Table 4). The rating scale (especially the first/starting point of the scale) also had challenges and posed threat to the accuracy of the data obtained.

3.3. Specific item measurement report

Table 5 presents the summary statistics for the item facet.

The model summary in Table 5 showed a significant chi-square test, $\chi^2(12) = 72.1, p < .001$. The outcome of the chi-square test suggests that the items were not of equal difficulty. The separation strata index of 3.09 implies that the items have 3 statistically distinct levels of difficulty. The separation ratio of 2.07 and reliability of .81 (values close to 0 are preferred) provided much evidence for the fact that the heterogeneity (in terms of item difficulty) among the items was more than expected.

As presented in Table 5, the item measurement report indicated that 5 out of 13 had acceptable fit indices indicating that they were clear and/or not redundant in the measurement of the constructs in the evaluation exercise (items 2, 7, 9, 12, and 11). Three of the remaining items were overfitting (items 1, 5, and 10) whereas five of them were misfitting (items 3, 4, 8, and 13). The items which recorded overfit meant that they were redundant; these items failed to provide new information to the measurement of the construct. Items that recorded misfit, on the other hand, implied that the items were unclear to the student-raters, and may not form part of the set of items that together define the single measurement construct.

3.4. Criteria measurement report

The criteria measurement report provides information on the role played by the sub-dimensions of the instrument. Two sub-dimensions

Table 3. Individual students' rating behaviours.

No.	Observed rating	Expected rating	Logit measure	S.E.	Infit		Outfit	
					MnSq	ZStd	MnSq	ZStd
12	2.62	2.77	2.64	.35	1.72	.90	2.05	1.2
41	2.31	2.64	2.60	.23	1.92	1.5	1.97	1.4
15	2.46	2.75	2.44	.30	1.82	1.6	1.81	1.4
6	2.92	3.10	2.88	.32	1.85	1.0	1.90	1.1
94	3.31	3.35	2.99	.31	1.85	.8	2.09	1.4
21	3.74	3.69	-2.68	.50	.71	-.21	.41	.60
128	3.85	3.34	-2.01	.78	.66	-.10	.59	-.40
110	3.79	3.67	-2.67	.67	.62	-.40	-.52	-.06
129	4.0	3.87	-3.07	1.85	.21	-.18	.11	-.14
69	4.0	3.88	-3.24	1.85	.16	-.13	.13	-.12

MnSq- Mean-square; ZStd-standardised infit/outfit statistics.

Table 4. Unexpected responses for the model.

Scale cat.	Observed score	Expected score	Residual	Std. residual	Item	Criteria
1	1	1.9	-.90	-4.20	8	MoD
3	3	3.9	-.90	-4.10	6	MoD
4	4	3.1	.90	4.0	5	MoD
1	1	2.8	-1.8	-3.70	1	MoD
1	1	3.6	-2.6	-3.70	8	MoD
1	1	2.8	-1.8	-3.60	3	MoD
1	1	2.7	-1.7	-3.30	1	MoD
1	1	3.4	-2.4	-3.30	6	MoD
1	1	2.7	-1.7	-3.20	6	MoD
1	1	2.7	-1.7	-3.0	9	MoD

MoD- Mode of delivery.

were the focus: mode of teaching and course contents. The report seeks to assess whether each of the scales functioned as whether the items under specific sub-dimension measuring the construct in question or sub-dimensions were redundant. This is a follow-up to the summary of the item report shown in Table 5 which found traces of criterion dysfunction. Table 6 presents the details of the results on the criteria.

The results on the mode of delivery criterion, as presented in Table 6, showed a significant chi-square test which tests the hypothesis that the items are homogeneous in terms of measuring the construct of interest, χ^2 (9) = 64.8, $p < .001$. Based on the test, it was revealed that the mode of delivery scale was psychometrically multi-dimensional. The scale was found to have three distinct sub-dimensions as shown by the separation strata index of 3.28 and a separation ratio of 2.21. The reliability of separation was .83 which showed heterogeneity among the items measuring the mode of teaching.

The results on the course content dimension revealed a non-significant chi-square test which tests the hypothesis that the items (under course content) are homogeneous in terms of measuring the construct of interest, χ^2 (2) = 3.4, $p = .180$. It was evident, from the results, that the course content scale was psychometrically unidimensional. This was also consistent with the separation strata index of 1 and the reliability of separation closer to zero ($R = .10$). Generally, it was easy for the lecturers to obtain a higher score under the mode of delivery dimension than the course content section.

3.5. Scale functioning quality

The quality of the scale used was also explored. For the items, different forms of scales were used for the sub-scales. The scales were, (1) Not very well -> Not well -> Well -> Very well, (2) Not detailed -> Slightly detailed -> Detailed -> Very detailed, (3) Not likely -> Slightly likely -> Likely -> Very likely, and (4) Less than 70% -> 70–79% -> 80–89% -> 90% or more. For each scale, the starting point is given a value of "1" which depicts that the trait being measured is rarely present. The highest point was given a value of "4" indicating that the trait being measured is highly present. The scales were examined to find out whether the raters were able to accurately use them and thus, the response options was functioning appropriately. To do this, some of the raters were sampled to study their use of the scale. Table 7 present the details of the scale functioning quality.

The results in Table 7 showed that some of the raters were not able to appropriately use the scale. Taking rater 6, for instance, scale category 1 recorded a very low frequency count and percentage ($n = 1$, 8%) indicating that the scale category was unclear for the rater and should be revised. The logit measure for the scale categories did not follow the expected pattern; scale category 1 had a logit measure of .83, category 2

Table 5. Summary statistics for item measurement report.

No.	Logit measure	S.E.	Infit		Outfit		Remark
			MnSq	ZStd	MnSq	ZStd	
1	-1.32	.16	.94	-.30	.87	-.90	**Overfit
2	-1.32	.16	.93	-.50	1.00	.0	Acceptable
3	-1.40	.16	1.25	1.7	1.15	.70	Misfit
4	1.35	.14	1.13	1.2	1.14	1.4	Misfit
5	1.25	.14	.95	-.4	1.16	1.5	**Overfit
6	-1.27	.16	1.26	1.9	1.14	1.0	Misfit
7	1.06	.15	.95	-.40	.92	-.06	Acceptable
8	-1.35	.16	1.16	1.1	1.00	.0	***Misfit
9	-1.30	.16	1.00	.10	.87	-.90	Acceptable
10	1.77	.14	.71	-3.5	.72	-3.2	Overfit
12	1.07	.15	.98	-.10	.92	-.60	Acceptable
11	1.07	.15	.98	-.10	.92	-.06	Acceptable
13	1.39	.14	1.29	2.6	1.22	2.2	Misfit
Model Summary				Estimate			
RMSE				.15			
Separation ratio				2.07			
Reliability of separation				.81			
Separation strata index				3.09			
Chi-squared				72.1			
Df				12			
p-value				.000*			

RMSE- Root Mean Square Error; S.E- Standard Error; SD- Standard Deviation; MnSq- Mean-square; ZStd-standardised infit/outfit statistics; *significant at p < .001; **Remark for outfit statistic only; ***Remark for infit statistic only.

Table 6. Criteria measurement report.

	Criteria/Sub-dimensions	
	Mode of Delivery	Course Content
Number of items	10	3
Logit measure	-1.05	1.18
RMSE	.15	.15
Adj. (True) S.D	.34	.05
Separation ratio	2.21	.33
Reliability of separation	.83	.10
Separation strata index	3.28	1
Chi-squared	64.8	3.4
df	9	2
p-value	.000*	.180

* significant at $p < .001$.

had a logit measure of .60, category 3 had .39 and category 4 had a logit measure of .51. It can be observed that instead of the average logit measure increasing along with the scale categories, it rather decreased. The MnSq outfit for all the categories was greater than 1 suggesting that the scales did not contribute to the meaningful measurement of the trait by rater 6.

Raters 43, 60, 154, 2108, and 2613 did not use the scale category 1 which showed that the category was quite unclear to them. Rater 43, for example, was found assigning lecturers with the same ability different scores. Lecturers with a proficiency of, say, -4 (logit measure) had the highest chance of getting a score of 2 instead of 1 indicating that the rater failed to discriminate between scale categories 1 and 2 (see Table 7).

3.6. How fair lecturers are rated in terms of their quality of teaching

The study further assessed the extent to which lecturers are fairly rated by the students regarding the appraisal of courses and teaching. The details of the results are shown in Tables 8, 9, and 10.

The result in Table 8 highlights the fit statistics for instructors/lecturers which provides information on whether the lecturers' abilities were well measured. The results showed that about 32.4% (infit) and 28.3% (outfit) had acceptable fit suggesting that they received ratings that were consistent with their actual abilities. The majority of the instructors were misfitting (47.6%) and overfitting (20%) suggesting that the abilities of the lecturers were not measured properly or the description of the course contents was far from what exists in reality.

Table 9 further provides a summary of the person (lecturer) measurement report which presents details of the entire model.

The results, shown in Table 9 offer insight into understanding the overall model. The results revealed a non-significant chi-square test, $\chi^2(144) = 6.1.1$, $p = .730$. The chi-square test sought to test the hypothesis that all the lecturers were equivalent in terms of their ability to teach and handle specific courses assigned to them. Based on the chi-square test, it was found that the lecturers were equivalent in terms of their ability to teach and handle courses. This was supported by the separation strata index of 1.0 (supposed to be greater than 1), reliability of separation of .01 (high separation reliability is required reliability), and the separation ratio (a high separation ratio is required). From all indications, there was low discrimination in terms of the ability of the objects of measurement (instructors). The location summary showed that there is an overestimation of the ability of lecturers with an observed average of 3.88 and a fair/expected average of 3.47. The logit measure of 2.67 reinforces the fact that there were more high scores than low scores. The overall infit and outfit statistics indicate that the lecturers' abilities were not accurately measured.

Some of the specific cases of poor fit (misfit) were selected and studied. This is presented in Table 10.

Table 7. Scale functioning quality.

Rater	Category	Counts	% Used	Quality Control		
				Logit measure	Expected measure	Outfit MnSq
Rater 6	1	1	8.0	.83	.27	1.80
	2	3	23.0	.60	.41	1.30
	3	5	38.0	.39	.53	1.80
	4	4	31.0	.51	.63	1.10
Rater 12	1	2	15.0	.12	-.42	1.90
	2	2	15.0	-.24	-.28	1.10
	3	8	62.0	-.33	-.16	1.50
	4	1	8.0	.13	-.07	.90
Rater 13	1	2	15.0	.76	.62	1.20
	2	–	–	–	–	–
	3	2	15.0	.84	.76	1.30
	4	9	69.0	.83	.87	1.10
Rater 18	1	1	8.0	.97	.84	1.10
	2	–	–	–	–	–
	3	5	38.0	.74	.98	.60
	4	7	54.0	1.24	1.09	.80
Rater 19	1	1	8.0	.91	1.06	.60
	2	1	8.0	.95	1.20	.40
	3	–	–	–	–	–
	4	11	1.37	1.37	1.33	1.0
Rater 22	1	1	8.0	.18	.76	.60
	2	–	–	–	–	–
	3	6	46.0	.80	.90	.70
	4	6	46.0	1.21	1.01	.80
Rater 43	1	–	–	–	–	–
	2	1	8.0	1.24	.67	1.4
	3	7	54.0	.76	.81	.90
	4	5	38.0	.87	.92	1.0
Rater 60	1	–	–	–	–	–
	2	1	8.0	1.24	.67	1.4
	3	7	54.0	.74	.81	.80
	4	5	38.0	.90	.92	1.0
Rater 92	1	1	8.0	.96	.40	2.1
	2	1	8.0	1.03	.54	2.1
	3	5	38.0	.36	.67	.30
	4	6	46.0	.86	.78	.90
Rater 154	1	–	–	–	–	–
	2	1	9.0	.36	.88	.40
	3	2	18.0	1.47	1.04	1.8
	4	8	73.0	1.15	1.19	1.0
Rater 448	1	1	8.0	.70	.91	.70
	2	–	–	–	–	–
	3	4	31.0	.92	1.04	.80
	4	8	62.0	1.25	1.16	.90
Rater 752	1	–	–	–	–	–
	2	2	15.0	.61	.33	1.30
	3	6	46.0	.40	.46	.90
	4	5	38.0	.53	.57	1.10
Rater 1149	1	1	8.0	.85	1.05	.70
	2	–	–	–	–	–
	3	2	15.0	.68	1.16	.30
	4	10	77.0	1.39	1.28	.80
Rater 1825	1	–	–	–	–	–
	2	2	15.0	.15	.08	1.00
	3	8	62.0	.20	.21	1.20
	4	3	23.0	.28	.31	1.00

(continued on next page)

Table 7 (*continued*)

Rater	Category	Counts	% Used	Quality Control		
				Logit measure	Expected measure	Outfit MnSq
Rater 2108	1	–	–	–	–	–
	2	1	8.0	.49	.67	.80
	3	7	54.0	.88	.81	1.20
	4	5	38.0	.85	.92	1.10
Rater 2613	1	–	–	–	–	–
	2	2	15.0	.38	.42	1.10
	3	5	38.0	.62	.55	1.10
	4	6	46.0	.63	.67	1.00

Table 8. Fit statistics for Instructors/Lecturers.

Fit Statistics	Indicators	Infit	Outfit
<.90	Overfit	29(20.0%)	31(21.4%)
.90–1.10	Acceptable fit	47(32.4%)	41(28.3%)
>1.10	Misfit/Underfit	69(47.6%)	73(50.3%)

Table 9. Summary statistics for person measurement report.

Model Summary	Estimate	Location Summary	Estimate
RMSE	.19	Observed Average ± SD	3.88 ± .12
Separation ratio	.01	Fair(M) Average ± SD	3.47 ± .13
Reliability of separation	.01	Logit Measure	2.65
Separation strata index	1.0	Model S.E.	.17
Chi-squared	6.1	Infit	2.01
df	144	Outfit	1.98
p-value	.730		

RMSE- Root Mean Square Error; S.E- Standard Error; SD- Standard Deviation.

As shown in Table 10, the logit measure of the selected cases ranged from 2.43 to 2.80 indicating a high ability measure. A closer look at the observed ability and expected ability estimates indicates that the lecturers were consistently rated higher. Lecturer 117, for instance, was given a score of 3.58 instead of an expected score of 3.28. Lecturer 008 was also rated with a score of 3.52 which deviated from his/her true rating of 3.41. Likewise, instructor 010 was rated with a score of 3.72 instead of 3.52.

4. Discussion

The results of this present study revealed that the items and raters played a very significant role in contributing to the variabilities in students' ratings. The results indicated that it was either the lecturers were differently rated by the students on different items or the variances were contributed by other random errors. The results from the MFRM analysis found that the majority of the evaluation items did not function as it was supposed to. About 61% of the items were flagged as problematic; some of these items were identified as unclear or failed to measure the trait being measured, and others were redundant in terms of measuring a significant aspect of the trait. Furthermore, the study discovered that the scale categories used for the rating significantly contributed to the errors of measurement in students' responses regarding the appraisal of courses and teaching. Consequently, the students were not able to use the scale categories as expected and this led to non-functional rating scales. Just like the results of this study, Börkan found that the rating scale (5-points) which was used for the survey did not function as expected. It was found that the items were non-equivalent in their level of difficulty.

This result signifies that students systematically differed in the way they evaluated the same lecturer. The result could mean that individual students evaluated the same lecturers based on what they think constitute teaching quality. Several previous studies corroborate this results (e.g., Feistauer and Richter 2016; Quansah, 2020; VanLeeuwen et al., 1999). Feistauer and Richter (2016), for example, found a high level of rater uncertainty and variability in students' evaluation of teaching exercise at the University of Kassel in Germany. In Quansah's (2020) study, rater inconsistencies were found as the second-largest source of variability in students' appraisal of teaching in a university in Ghana. Similarly, VanLeeuwen et al. (1999) also found high rater variability in students' appraisal of instruction. For all these studies, although rater variability was found, the nature of this variability was not explored. As to whether raters were lenient or severe and what is causing this variability. In this present study, however, MFRM analysis was conducted to do this.

The results showed that the majority of the students were lenient when appraising courses and teaching quality. This could be attributed to the notion that the students were afraid/uncomfortable to rate a lecturer poorly due to some reasons like being in the good books of the lecturer. Whereas some students were not able to discriminate between different performance levels, others were prone to the halo effect- ratings that are influenced by other factors other than the specific behaviours being measured. There are specific instances where students will be caught in the web of the halo effect. These may include situations where students provide ratings based on a less salient dimension or construct. For example, a student may be impressed by a lecturer's punctuality to class and based on this behaviour may rate the lecturer excellently on his/her teaching quality. In this case, a positive impression created by the lecturer have influenced the students when rating the lecturer on a different dimension. In some instances, a student may have a negative impression of a lecturer (based on a specific behaviour exhibited); this

Table 10. Sample of instructor ability measure statistics.

No.	Observed Ability	Expected Ability	Ability Measure (Logit)	S.E.	Infit		Outfit	
					MnSq	ZStd	MnSq	ZStd
117	3.58	3.28	2.80	.45	1.90	1.10	1.87	1.20
001	3.65	3.44	2.77	.35	1.91	1.60	1.86	1.80
013	3.86	3.64	2.74	.35	1.98	1.30	1.98	1.10
008	3.52	3.41	2.68	.31	1.40	1.20	1.96	1.10
050	3.73	3.55	2.66	.12	1.99	1.10	1.97	1.30
066	3.70	3.47	2.66	.21	1.70	1.70	1.40	1.30
119	3.79	3.57	2.62	.13	1.94	1.70	1.91	1.1
010	3.72	3.52	2.57	.12	1.40	1.50	1.25	1.50
004	3.79	3.58	2.57	.13	1.30	1.30	1.97	1.30
134	3.42	3.23	2.43	.10	1.18	1.90	1.15	1.4

MnSq- Mean-square; ZStd-standardised infit/outfit statistics.

can also lead to the student providing a low rating for the lecturer. It is not always so that the lecturer has created a positive or negative impression for the students, but some raters (students) may just fail to distinguish between the exact behaviours being measured and other behaviours of the lecturer unrelated to the tagreted behaviours. These behaviours of students which constituted the halo effect were also found among university students in Börkan's (2017) study. The consistency in the results of this study and Börkan's research speaks to the fact that university students, irrespective of their location, have similar behaviours when rating their lecturers.

High variability was found among raters such that the students were not similar in their level of severity when rating. This result agrees with the view of Eckes (2015) who argued that rather than operating on collective grounds, student-raters regularly seem to considerably differ regarding deeply fixed, more or less individualised rating predispositions and thereby threatening the validity of the evaluation outcomes. A similar result was also found in a study conducted by Börkan (2017) who also used MFRM to explore the sources of variations in students' appraisal of teaching in a university in Turkey. In Börkan's study, it was also revealed that a larger proportion of the students rated leniently. It appeared that the issue of rater variability in students' evaluation of teaching is viewed differently in studies that used CMT. Almost all studies that used the CMT approach to investigate rater variability found a relatively high level of consistency among the raters. For instance, research conducted by Zhang et al. (2015) to investigate rater consistency among students in student evaluation exercise in two universities in the US, found that the students were consistent with their ratings regardless of the qualities they sought after in an instructor. In another study conducted by Fah and Osman (2011), a high level of consistency among students in their ratings of instructors was confirmed. Other studies by Raza and Irfan (2018), and Samian and Noor (2012) had results that supported the presence of homogeneity of ratings among raters. Ironically, the results from these studies which employed CMT are not comprehensive as they focus on only one source of measurement error at a time. The results of these studies only operate on the assumption of stability of student ratings over time, neglecting measurement errors which can be due to item heterogeneity, non-functional rating scale, and the interaction of several main effect variables. It is not surprising that these studies found some level of consistency in students' ratings of instructors.

4.1. Practical implications

The findings of this research studied and combined several forms of validity evidence to support the use and interpretation of the data obtained from students regarding their evaluation of teaching quality. Consequently, it should be emphasized that the various measurement reports from this study should not be examined independently but as a whole. For example, although the study showed that the lecturers' teaching abilities were homogeneous, this finding is not conclusive; this is because other pieces of evidence suggest an overestimation of the lecturers' ratings provided by the students. Hence, the homogeneity of teaching ability could be attributed to the inacurate ratings from the students. There is, therefore, the need to combine the various pieces of information before conclusions are drawn.

The outcome of the analysis highlights the lapses in the appraisal exercise from the perspectives of students' rating behaviours, items/scale functioning, and rating fairness. The findings have implications for the utility of the evaluation results in terms of making high-stakes decisions such as promotion. Essentially, appraisal outcomes may have positive and/or negative consequences for lecturers/teachers who are the evaluation objects. For example, a teacher who is poorly rated by severe raters for being strict may be denied promotion and tenure renewal. Meanwhile, a teacher who has instructional pedagogical or delivery problems may be rated by lenient raters and might be promoted. Considering the variability surrounding this appraisal exercise, lecturers/teachers may put up some behaviours (unrelated to instructional delivery/mode of delivery) to please students to obtain excellent ratings. There is a need for rethinking and redesigning students' appraisal exercises that limit biases and eliminate spurious variables from the point of designing, implementing, and utilising the appraisal results. University administrators should explore avenues to improve students' behaviour and attitude towards the exercise and redesign high-quality questions/items/instruments to promote rating fairness for lecturers for the right decisions to be made.

4.2. Limitations

Despite the significance of the study, there were some limitations. First, the data used were for a single semester only (i.e., the second semester of the 2019/2020 academic year) and thus, the results may not be representative enough for generalisation. Further, factors such as gender distribution, lack of demographic variables for the data, replication of responses, and biases in course representation were not controlled by the investigator because these variables were not available in the dataset due to ethical reasons. The outcome of this research may be limited in terms of its applicability to other institutions of higher education due to differences in the evaluation of teaching exercises. Accordingly, different educational institutions may have different evaluation instruments and items, and may even attempt to measure traits of teaching and activities differently. Although the findings serve as a prompt for other institutions in Ghana and beyond, it would be difficult to apply the findings of this research to other higher education enterprises.

4.3. Conclusions and recommendations

The study concludes that the quality of data provided by students in relation to evaluating courses and teaching has validity concerns. The non-functional structure of items, less effective functioning of rating scales, and inconsistent rating behaviours of students cast doubts on the validity of students' appraisal of courses and teaching data. Based on this evidence, the dependability of the student evaluation of courses and teaching can be considered relatively low. It is recommended that the management of UCC and DAPQA-UCC should train students on how to rate accurately to reduce errors of measurement emanating from students (such as halo effect, inconsistent rating, and inability to use the rating scales) during the appraisal of courses and teaching. The training should include what behaviours should be looked out for when appraising, what constitutes excellent, average or high performance, and how to reduce the effect of extraneous variables (such as the friendliness of the lecturer) from influencing their rating. The study recommends that DAPQA-UCC should revisit the items on the existing evaluation form and the items should be further subjected to rigorous validation. This is to make the items more functional in terms of clarity and construct validity. Besides, the scale categories used for the rating exercise should be changed or modified by the directorate (DAPQA-UCC). Particularly, scale category descriptors should be provided on the evaluation form to guide the ratings of the students.

Declarations

Author contribution statement

Frank Quansah, PhD: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

The work was supported by the School of Graduate Studies and Graduate Students Association (GRASSAG) of the University of Cape Coast, through the Samuel and Emelia Brew-Butler/SGS/GRASSAG-UCC Research Grant.

Data availability statement

The author do not have permission to share data.

Declaration of interest's statement

The author declare no competing interests.

Additional information

No additional information is available for this paper.

Acknowledgements

None declared.

References

Adams, M.J., Umbach, P.D., 2012. Nonresponse and online student evaluations of teaching: understanding the influence of salience, fatigue, and academic environments. Res. High. Educ. 53 (5), 576–591.

Alter, M., Reback, R., 2014. True for your school? How changing reputations alter demand for selective U.S. colleges. Educ. Eval. Pol. Anal. 36 (3), 346–370.

Becker, W.E., Bosshardt, W., Watts, M., 2011. Revisiting How Departments of Economics Evaluate Teaching. Technical Report, Working Paper Presented at the Annual Meetings of the American Economic Association.

Börkan, B., 2017. Exploring variability sources in student evaluation of teaching via many-facet Rasch model. J. Measur. Eval. Educ. Psychol. 8 (1), 15–33.

Braga, M., Paccagnella, M., Pellizzari, M., 2011. Evaluating students' evaluations of professors. Inst. Study Labor 5620, 1–54.

Eckes, T., 2015. Introduction to many-facet Rasch Measurement: Analysing and Evaluating Rater-Mediated Assessment, second ed. Peter Lang GmbH, Frankfurt.

Ewing, A.M., 2012. Estimating the impact of relative expected grade on student evaluations of teachers. Econ. Educ. Rev. 31 (1), 141–154.

Fah, B.C.Y., Osman, S., 2011. A case study of student evaluation of teaching in university. Int. Educ. Stud. 4 (1), 44–50.

Feistauer, D., Richter, T., 2016. How reliable are students' evaluations of teaching quality? A variance components approach. Assess Eval. High Educ. 10, 1–17.

Galbraith, C.S., Merrill, G.B., Kline, D.M., 2012. Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business-related classes? A neural network and Bayesian analyses. Res. High. Educ. 53 (3), 353–374.

Goffin, R.D., Olson, J.M., 2011. Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. Perspect. Psychol. Sci. 6 (1), 48–60.

Goos, M., Salomons, A., 2017. Measuring teaching quality in higher education: assessing selection bias in course evaluations. Res. High. Educ. 58 (1), 341–364.

Gyimah, E.K., Kwarteng, A.J., Anane, E., Nkrumah, I.K., 2016. Lecturers' perception of students' appraisal of courses and teaching: a case of University of Cape Coast, Ghana. Int. J. Res. Comm. IT Manag. 6 (9), 21–26.

Hornstein, H.A., 2017. Student evaluations of teaching are inadequate assessment tool for evaluating faculty performance. Cogent Educ. 4 (1), 13–42.

Jol, G., Stommel, W., 2016. Ethical considerations of secondary data use what about informed consent? Dutch J. Appl. Linguist. 5 (2), 180–195.

Ko, J., Sammons, P., Bakkum, L., 2013. Effective Teaching: A Review of Research and Evidence. CfBT Education Trust, Berkshire.

Kwarteng, A.J., Doku, D.T., Matta, D.A.P., Doh-fia, S., 2014. University involvement in the university quality assurance systems (Qas): a case of University of Cape Coast, Ghana. Researchjournali's J. Manag. 2 (8), 1–15.

Kirkpatrick, D.L., 1959. Techniques for evaluating training programs. J. Am. Soc. Train. Direct. 13, 3–9.

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

Linacre, J.M., 1989. Many-*Facet Rasch Measurement*. MESA Press, Chicago.

Linacre, J.M., 1994. Many-Facet Rasch Measurement, second ed. MESA Press, Chicago.

Linacre, J.M., 2003. A User's Guide to FACETS [computer Program Manual]. MESA Press, Chicago.

Linacre, J.M., Wright, B.D., 2002. Understanding Rasch measurement: Construction of measures from many-facet data. J. Appl. Meas. 3 (4), 486–512.

Lunz, M., Wright, B., Linacre, J., 1990. Measuring the impact of judge severity on examination scores. Appl. Meas. Educ. 3 (4), 331–345.

MacNell, L., Driscoll, A., Hunt, A., 2015. What's in a name: exposing gender bias in student ratings of teaching. Innovat. High. Educ. 40 (4), 291–303.

Masters, G.N., 2010. The partial credit model. In: Nering, M.L., Ostini, R. (Eds.), Handbook of Polytomous Item Response Theory Models. Routledge, New York, NY, pp. 109–122.

McPherson, M.A., Jewell, R., 2007. Levelling the playing field: should student evaluation scores be adjusted? Soc. Sci. Q. 88 (3), 868–881.

McPherson, M.A., Jewell, R.T., Kim, M., 2009. What determines student evaluation scores? A random-effects analysis of undergraduate economics classes. E. Econ. J. 35 (1), 37–51.

Ogbonnaya, U.I., 2019. The reliability of students' evaluation of teaching at secondary school level. Probl. Educ. 21st Century 77 (1), 97–109.

Quansah, F., 2020. An assessment of lecturers' teaching using generalisability theory: a case study of a selected university in Ghana. S. Afr. J. High Educ. 34 (5), 136–150.

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. Assess Eval. High Educ. 38, 224–239.

Raza, S.A., Irfan, M., 2018. Students' evaluation of teacher attributes: implications for quality in higher education. Bull. Educ. Res. 40 (1), 197–214.

Samian, Y., Noor, N.M., 2012. Students' perception of good lecturer based on lecturer performance assessment. Proc. Soc. Behav. Sci. 56 (1), 783–790.

Sliusarenko, T., 2013. Quantitative Assessment of Course Evaluations. Technical University of Denmark, Kgs Lynby.

Spooren, P., Brockx, B., Mortelmans, D., 2013. On the validity of student evaluation of teaching: the state of the art. Rev. Educ. Res. 83 (4), 598–642.

Steele, L.M., Mulhearn, T.J., Medeiros, K.E., Watts, L.L., Connelly, S., Mumford, M.D., 2016. How do we know what works? A review and critique of current practices in ethics training evaluation. Account. Res. 23 (6), 319–350.

Taut, S., Rakoczy, K., 2016. Observing instructional quality in the context of school evaluation. Learn. InStruct. 46, 45–60.

Theall, M., Franklin, J.L., 2001. Looking for bias in all the wrong places: a search for truth or a witch hunt in student ratings of instruction? N. Dir. Inst. Res. 109, 45–56.

University of Cape Coast, UCC, 2012. Directorate of Academic Planning and Quality Assurance: Vision, mission and Core Responsibilities. Retrieved from. http://ucc.edu.gh/apqa/sites/ucc.edu.gh.apqa/files/dapqa/strate.gic/plan.pdf. (Accessed 23 February 2020). accessed.

University of Cape Coast, UCC, 2015. Criteria for Appointment and Promotion. University Press, Cape Coast.

VanLeeuwen, D.M., Dormody, T.J., Seevers, B.S., 1999. Assessing the reliability of student evaluation of teaching (SET) with generalisability theory. J. Agric. Educ. 40 (4), 1–9.

Wilhelm, W.B., 2004. The relative influence of published teaching evaluations and other instructor attributes on course choice. J. Market. Educ. 26 (1), 17–30.

Wright, R., 1996. A study of the acquisition of verbs of motion by Grade 4/5 early French immersion students. Can. Mod. Lang. Rev. 53 (1), 257–280.

Wu, M., Adams, R.J., 2013. Properties of Rasch residual fit statistics. J. Appl. Meas. 14, 339–355.

Zhang, S., Fike, D., DeJesus, G., 2015. Qualities university students seek in a teacher. J. Econ. Econ. Educ. Res. 16 (1), 42–54.