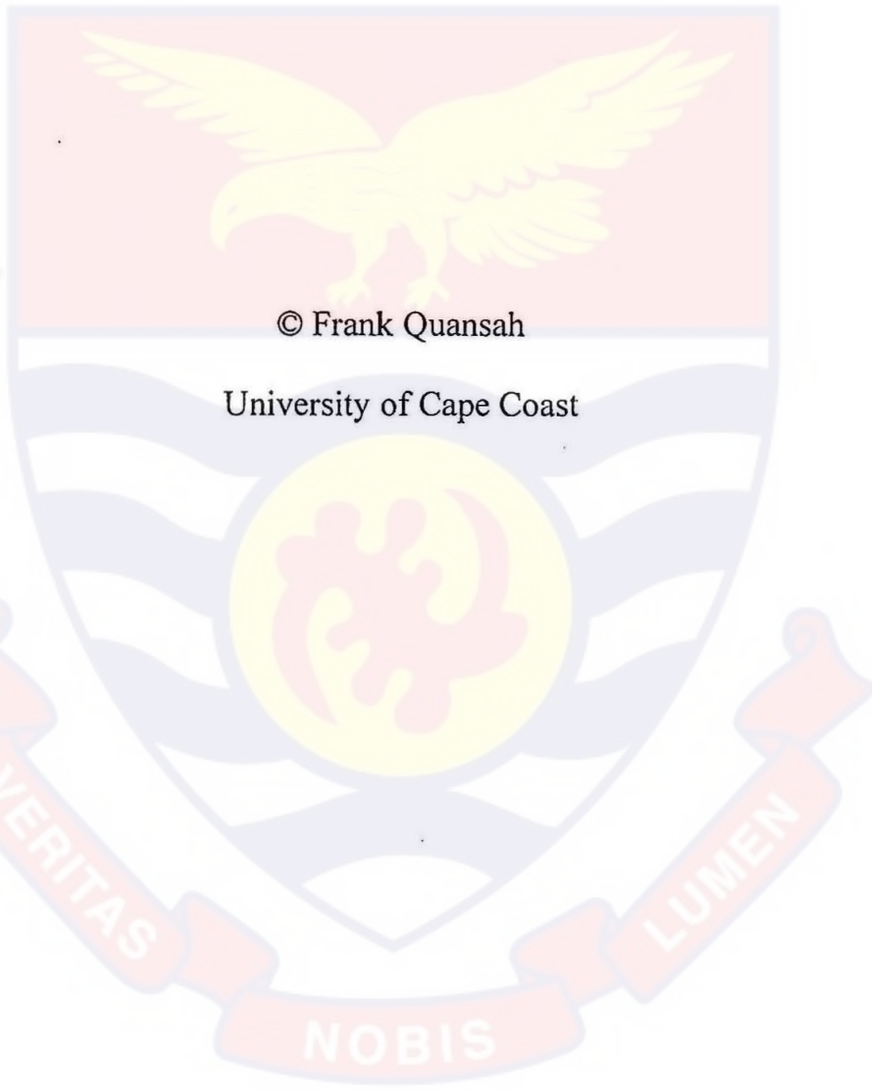
© University of Cape Coast https://ir.ucc.edu.gh/xmlui



© Frank Quansah

University of Cape Coast

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

UNIVERSITY OF CAPE COAST

VALIDITY OF STUDENTS' EVALUATION OF COURSES AND TEACHING USING GENERALIZABILITY THEORY AND MANY-FACET RASCH MODELLING IN THE UNIVERSITY OF CAPE COAST, GHANA

BY

FRANK QUANSAH

Thesis submitted to the Department of Education and Psychology of the Faculty of Educational Foundations, College of Education Studies, University of Cape Coast, in partial fulfillment of the requirements for the award of Doctor of Philosophy degree in Measurement and Evaluation.

CALL No.
ACCESSION No.
7314
CAT. CHECKED
FINAL CHECKED

JULY 2021

SAM JONAH LIBRARY
UNIVERSITY OF CAPE COAST
CAPE COAST

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

DECLARATION

Candidate's Declaration

I hereby declare that this thesis is the result of my own original research and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature: ..... Date: 20TH JUNE, 2022

Name: FRANK QUAYSON

Supervisors' Declaration

We hereby declare that the preparation and presentation of the thesis were supervised in accordance with the guidelines on supervision of thesis laid down by the University of Cape Coast.

Principal Supervisor's Signature: ..... Date: 20-06-22

Name: DR KENNETH ASAMOAH-GYIMAH

Co-Supervisor's Signature: ..... Date: 20-06-22

Name: DR. ANDREWS COBBINAH

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

ABSTRACT

The overarching purpose of this research was to examine the validity of students' evaluation of teaching and courses in the University of Cape Coast (UCC) through the lenses of Generalizability Theory and the Many-Facet Rasch Model. The study was grounded in the positivists' paradigm using the hybrid quantitative approach. Two main quantitative designs were used: a two-facets partially nested random unbalanced design and the three-facets design with a partial credit model. The study covered students for the 2019/2020 academic year in UCC. The study used secondary data on students obtained from the Directorate of Academic Planning and Quality Assurance, UCC (DAPQA-UCC). All cases and data points were involved in this study with a sample size of 24,726 regular students. The data were analysed by conducting univariate GT and partial credit MFRM analyses. The study revealed a low to moderate level of dependability of students' appraisal of courses and teaching and as such, lecturers were unfairly rated. It was further found that the sources of measurement error in the student evaluation exercise were 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 sensitise and train students on the evaluation of teaching, and as well review the existing evaluation form for appraising courses and teaching by subjecting the instrument to rigorous validation procedures.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

KEYWORDS

Course appraisal

Generalizability theory

Many-facet Rasch modelling

Mode of teaching

Reliability

Student evaluation

Validity

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

ACKNOWLEDGEMENTS

I wish to express my profound gratitude to my supervisors, Dr. Kenneth Asamoah-Gyimah and Dr. Andrews Cobbinah, all from the Department of Education and Psychology, University of Cape Coast, for their timely feedback, responsiveness, guidance and support throughout the conduct of this work. I also express my appreciation to Prof. Emmanuel Kofi Gyimah, the Acting Director, Directorate of Academic Planning and Quality Assurance, University of Cape Coast (DAPQA-UCC), for welcoming the idea and granted me access to the data.

I am thankful to Dr. Regina Mawusi Nugba for her encouragement, support, and advice she gave me. I am grateful also to Dr. Prince Asare and Dr. Gabriel Kwasi Aboagye, who helped in proofreading the work at the early stages of the thesis. I am indebted to my lovely wife, Mrs. Joyce Quansah, Miss. Shalleen Essilfie, Dr. Vera Rosemary Ankoma-Sey and Dr. Gabriel Essilfie for their support, care and inspiration during the times of preparing this thesis report. I express my deepest appreciation to Mr. Kofi Ahenekwa, Pastor George Ofori, and all the executives of Shekinah Local of the Church of Pentecost (CoP) for their spiritual support throughout my programme. I am thankful also to my siblings (Lydia, Nathaniel, Ebenezer and Albert) for their reassurance during the thesis writing.

I, finally, wish to thank the School of Graduate Studies and Graduate Students Association (GRASSAG) all of the University of Cape Coast for awarding me the Samuel and Emelia Brew-Butler/SGS/GRASSAG-UCC Research Grant to support me with funds to conduct this research. I am much grateful for this opportunity.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

DEDICATION

To my late mother, Mrs. Elizabeth Quansah; my wife, Mrs. Joyce Quansah; my child, Gideon Quansah; and my father, Mr. Nana Kumi Quansah.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

TABLE OF CONTENTS

	Page
DECLARATION	ii
ABSTRACT	iii
KEYWORDS	iv
ACKNOWLEDGEMENTS	v
DEDICATION	vi
TABLE OF CONTENTS	vii
LIST OF TABLES	x
LIST OF FIGURES	xii
LIST OF ACRONYMS	xiii
CHAPTER ONE: INTRODUCTION	
Background to the Study	1
Statement of the Problem	17
Purpose of the Study	23
Research Questions	24
Significance of the Study	25
Delimitations	27
Limitations	27
Definition of Terms	29
Organisation of the Rest of the Study	29
CHAPTER TWO: LITERATURE REVIEW	
Brief Description of Literature Search Process	31
Review of Key Concepts	33
Theoretical Framework	41

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Differentiating between GT and MFRM 49

Identical Concepts Used in GT and MFRM 51

Theoretical Review 55

Empirical Review 70

Conceptual Framework 92

Chapter Summary 93

CHAPTER THREE: RESEARCH METHODS

Research Paradigm 95

Research Approach 96

Design for Generalizability Theory 96

Design for Many-Facet Rasch Modelling 100

Study Area 102

Population 104

Sampling Procedure 105

Data Collection Instrument 106

Data Collection Procedure 108

Ethical Consideration and Data Management 109

Data Processing and Analysis 110

Chapter Summary 123

CHAPTER FOUR: RESULTS AND DISCUSSION

Research Question 1 125

Research Question 2 127

Research Question 3 129

Research Question 4 135

Research Question 5 145

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Research Question 6 148

Research Question 7 169

Research Question 8 172

How the MFRM Complements GT in the Context of this Study 173

Discussion 175

Chapter Summary 185

CHAPTER FIVE: SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Overview of the Study 188

Key Findings 189

Contribution to Knowledge 192

Conclusions 194

Recommendations for Policy and Practice 195

Suggestions for Further Studies 197

REFERENCES 199

APPENDICES 229

APPENDIX A: SAMPLE OF EVALUATION FORM 230

APPENDIX B: INTRODUCTORY LETTER 235

APPENDIX C: ETHICAL CLEARANCE 236

APPENDIX D: LETTER FOR DATA ACCESS AND USE 237

APPENDIX E: SUPPORT LETTER FROM SUPERVISOR FOR DATA ACCESS 238

APPENDIX F: SAMPLES OF PROBABILITY CURVES 239

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

LIST OF TABLES

Table		Page
1	Summary of Empirical Studies	86
2	Two-Facet, Partially Nested Random Unbalanced G-study of Students' Appraisal of Lecturers	100
3	Details of Courses Assessed for 2019/2020 Second Semester	105
4	Summary of Analytical Procedure	121
5	Sources of Variability and their Variance Components for Course and Teaching Appraisal	126
6	Reliability Estimate of Student Appraisal of Courses and Teaching	128
7	Optimisation for Course Appraisal for Large Class Size	129
8	Optimisation for Course Appraisal Data for Small Class Size	130
9	Optimisation for Teaching Evaluation Data for Large Class Size	132
10	Optimisation for Teaching Evaluation with Small Class Size	133
11	Descriptive Statistics for Expected and Received Responses	136
12	Response Ratio of Some Sample Courses	138
13	Contribution of Non-response to Course Appraisal (n_i=3)	139
14	Influence of Non-response on Teaching Evaluation Data (n_i=10)	142
15	Fit Statistics for Students' Rating Behaviours	145
16	Summary Statistics for Rater (Student) Measurement Report	146
17	Individual Students' Rating Behaviours	147
18	Unexpected Responses for the Model	149
19	Summary Statistics for Item Measurement Report	150
20	Criteria Measurement Report	158
21	Scale Functioning Quality	159

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

22	Fit Statistics for Instructors/Lecturers	169
23	Summary Statistics for Person Measurement Report	169
24	Sample of Instructor Ability Measure Statistics	171
25	Bias Interaction for Course and Teaching Appraisal	172

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

LIST OF FIGURES

Figure		Page
1	Parents and Conceptual Framework of GT	44
2	Kirkpatrick Evaluation Model by Kirkpatrick (1959)	64
3	Conceptual Framework	92
4	Sources of Variability for (**r: p**) x **i** Design	99
5	Optimisation (Course) Graph with 20-items	131
6	Optimisation (Teaching) Graph with 12-items and (1-2) raters	134
7	Non-response graph for levels of study	137
8	Trend Analysis of Frequency of Non-response and Relative Error Variance for Course Appraisal	140
9	Trend Analysis of Number of Raters and G/Phi Coefficient for Course Appraisal	141
10	Trend Analysis of Frequency of Non-response and Relative Error Variance for Appraisal of Teaching	143
11	Trend Analysis of Number of Raters and G/Phi Coefficient for Appraisal of Teaching	144
12	Probability Curves for Raters 12 and 13	163
13	Probability Curves for Raters 43 and 60	165
14	Probability Curves for Raters 92 and 154	166
15	Probability Curves for Raters 2108 and 2613	168

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

LIST OF ACRONYMS

AQAU	Academic Quality Assurance Unit
CMT	Classical Measurement Theory
DAPQA	Directorate of Academic Planning and Quality Assurance
D-study	Decision study
G-study	Generalizability study
GT	Generalizability Theory
MFRM	Many-Facet Rasch Modelling
MnSq	Mean Square
PCM	Partial Credit Model
RSM	Rating Scale Model
UCC	University of Cape Coast

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

CHAPTER ONE

INTRODUCTION

In higher education, students play a critical role in the assessment of the quality of the educational services they provide (Amoako & Asamoah-Gyimah, 2020). It is not surprising that university students are allowed to evaluate the effectiveness of teaching and learning activities in the university. The commonest means is students' appraisal of courses, teaching, and learning experiences (Becker, Bosshardt, & Watts, 2011). The University of Cape Coast (UCC), Ghana, is no exception; Students are required to appraise courses and quality of teaching near the end of the semester through some online means. In this study, the researcher sought to examine the extent of dependability of ratings provided by students and also to understand the variables contributing to the variabilities in these ratings.

Background to the Study

The Sustainable Development Goal (SDG) four highlights the essentials for ensuring an inclusive and equitable quality of education, and stimulating lifetime learning opportunities for people from diverse backgrounds. Thus, the quality of education matters for achievements in learning (Hanushek, Kain, O'Brien, & Rivkin, 2005) and lifelong learning outcomes (Chetty, Friedman, & Rockoff, 2014b). Given this, many institutions of higher education, in Ghana and beyond, are making several efforts in the training of students for the business world (Quansah, Appiah, & Ankoma-Sey, 2019).

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

The universities' target of training skilled and competent labours fit for the knowledge economy has not been completely achieved (Quansah, Ankoma-Sey, & Asamoah, 2019), and this has created a cumulative pressure on institutions of higher education to attaining quality education, particularly in the teaching circles (Salem, 2014). While stakeholders (such as students, parents, and industry players) are also demanding a high level of quality education, university administrators are also continuously desiring approaches to improving the quality of education (Feistauer & Richter, 2016; Staufenbiel, Seppelfricke, & Rickers, 2016).

A significant component of quality education is instructional quality (Hanushek & Rivkin, 2006; Rivkin, Hanushek, & Kain, 2005). This notwithstanding, a consensus among scholars on the suitable approach to the measurement of quality of instruction is rarely available, especially, within the context of higher education where there are no standardised means of quantifying and comparing the instructional quality of lecturers within the same institution and among universities (Goos & Salomons, 2017). For the basic and pre-tertiary level of education, there has been consensus on the measurement of quality of teaching by regressing teacher fixed effects on the achievement of learners (Chetty, Friedman, & Rockoff, 2014a, Chetty et al., 2014b). This approach is inapplicable in higher education due to the unstandardized nature of performance indicators and courses taught across the various institutions; this has been extensively discussed in the literature (Cunha & Miller, 2014). With this, there is still a growing interest by the general public in the measurement of instructor and instructional quality in higher education institution (Goos & Salomons, 2017).

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Due to the intense competition existing among higher education institutions, universities today are undergoing a paradigm shift in students' status (Raza, Majid, & Zia, 2010) such that students are now freely making decisions regarding the type of institution to attend, the programme to choose, and even the type of major courses to read, just like customers selecting their preferred commodities in a supermarket with several varieties of product available (Raza & Khawaja, 2013). Hence, universities are expected to employ instructors and staff who would demonstrate the most desired qualities (Raza, 2014) which are capable of attracting and retaining consumers (i.e., students) (Douglas, McClelland, & Davies, 2008; Gruber, Reppel, & Voss, 2010; Helgesen & Nesset, 2007). Given the significant role students play in the sustenance and running of higher education institutions, they are referred to as stakeholders and as such expected to have a say in the affairs of the educational institution.

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; Zabaleta, 2007). Also, little evidence has been established that certified instructors are more efficacious than uncertified ones (Angrist & Guryan, 2008; Carrell & West, 2010; Kane, Rockoff, & Staiger, 2008). 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

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

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 promotion of instructors (Galbraith, Merrill, & Kline, 2012; Kogan, Schoenfeld-Tacher, & Hellyer, 2010), students' university application (Alter & Reback, 2014), and student's choice of courses (Wilhelm, 2004). In some advanced 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 sparkled extensive scientific literature in areas such as psychology, education, economics and sociology (Goos & Salomons, 2017). With this, 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 & Rakoczy, 2016).

The validity of students' evaluation of courses and teaching is a contentious issue (Hornstein, 2017). In an extensive review by Spooren, Brockx, and Mortelmans (2013), it was established that there are conceptual, theoretical and empirical supports for students' appraisal of courses and teaching. Despite this, the utilisation of data from students' evaluation of teaching as a proxy for 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 is polluted by noise. This presupposes that students have been found to evaluate instructional quality based on the characteristics of course (e.g., difficult/easy nature of courses), student (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

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

quality of teaching and course contents (Berk, 2005; Isely & Singh, 2005; Ko, Sammons, & Bakkum, 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 effect of students' evaluation of teaching and student scores, regardless of the learning outcomes (Ewing, 2012; Langbein, 2008; McPherson, 2006).

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, Driscoll, & Hunt, 2015; McPherson, Jewell, & Kim, 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 adjustment in the data to control for any extraneous variables (McPherson, 2006; McPherson & 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 was demonstrated 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.

In the view of Felder and Brent (2004), students' evaluation of teaching quality should be an integral part of instructional evaluation since:

> *"Students are in a better position than anyone else to assess certain components of courses and instruction, such as how interesting,*

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

> *clear, and fair a course facilitator is, and are consequently the only individuals who can say how a teacher has influenced their attitude towards the course subject, their motivation to learn it, and their self-confidence"* (p. 200).

Likewise, Prebble, Hargreaves, Leach, Naidoo, Suddaby, and Zepke (2004) indicated that students' appraisal of courses and teaching is one of the most valid and accessible indicators in the measurement of the quality of course content and teaching quality. This is because students can provide a more comprehensive and detailed rating of the educational and classroom practices of instructors as compared to using other procedures like self-evaluation, peer assessment or even appraisal by superior (Benton & Ryalls, 2016; Ogbonnaya, 2019; Prebble et al., 2004). Data from students' evaluation of teaching can offer significant information about instructors since the students spend most of their time with them (Quaglia & Corso, 2014).

Validity and sources of variability in students' rating of teaching

Validity is the soundness of the interpretation and uses of assessment or evaluation results; it is a matter of degree and all about bringing together pieces of evidence (Nitko & Brookhart, 2014). Applied to the context of this study, data from students' appraisal of teaching quality is highly valid when several pieces of evidence can be provided with regards to how the data were taken, the meaningfulness of the data and how the data were used (Nitko & Brookhart, 2014). For validity to be understood in the context of student evaluation, these possibilities should be looked at: (1) raters may not be accurate in their ratings such that the scores given may not reflects the skills and abilities of the instructors or overall teaching quality; (2) the evaluation items may not be clear

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

enough for raters to understand; and (3) students may rate other characteristics of the course and instructor other than actual psychological construct which should be rated. These issues, among others, are likely to influence the fairness of the evaluation exercise and the results thereof may not be a true reflection of the construct being measured.

In practice, when students are requested to evaluate course contents and teaching quality, there is a higher likelihood of disagreement among themselves due to many systematic and unsystematic factors (Eckes, 2015). The systematic factors include rater differences and for that matter differences in rater behaviours, the difficulty of items, and the central tendency effect. On the other hand, the random (unsystematic) factors comprise variations in physical scoring or testing conditions, attention fluctuations of raters as a result of fatigue, errors in transcription, and several others (Brennan, 2011). Whereas these systematic factors produce systematic errors, the unsystematic factors produce random errors (Eckes, 2015).

As a matter of fact, random (unsystematic) errors can easily be corrected through carefully planned assessment procedures. The case happens to be different when systematic errors are present. Systematic errors, unlike random errors, can be easily identified in a data set. It is worthy to note that random errors cancel out in large sample sizes and thus, do not usually have an effect on the validity of large data set (Eckes, 2015). Systematic errors, on the other side, create a pattern in the data set and distort the meaning derived from the data. Due to the critical nature of systematic errors and their implications on validity, this study focuses on identifying systematic errors in students' ratings

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

of quality of course contents and teaching and how these errors contribute to the variances in the ratings obtained from the students.

In most evaluation situations, giving a score to depict the degree to which a particular trait is possessed by the object of measurement (i.e., lecturers) is purely based on the subjective judgement of the rater (i.e., student). 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 (Eckes, 2015). Raters also have the probability of giving similar scores for a particular lecturer on a theoretically different criterion (Barrett, 2005; Iramaneerat & Yudkowsky, 2007). Due to this, several raters (as in the case of students' appraisal of instruction), in most cases, are utilised to cancel out random rater errors in the data obtained (Houston & Myford, 2009). The intent of using numerous raters is to have an estimate for every lecturer's instructional ability which is independent of some specific attribute of the raters. This is done by measuring the attributes of the raters and utilising such data to eliminate the faults of individual raters from the final score of the rater (Linacre, 1994).

Furthermore, raters usually carry out their appraisal of courses and teaching by using items with rating scales where points on the scale are required to signify, a consecutively higher degree of performance on the construct (i.e. teaching). Taking the analytic rating approach, for instance, raters lookout for specific features of the construct of interest and a score is assigned to that particular feature present. In the case of using the holistic rating method, raters evaluate the whole performance and a single score is assigned to performance (Engelhard, 2011). With each scoring type, raters are required to differentiate

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

between scale points and assign a rating that appropriately matches the performance; if this is not properly done, the validity of the score assigned will be threatened (Eckes, 2015). This, in most cases, could result in the over-utilisation of the middle point score category (for 5, 7, and 9 point-scale types) with few ratings at the two extreme ends of the scale. Overusing the middle point score results in the over-estimation of proficiency for low performers and under-estimation for high performers (Myford & Wolfe, 2003). This can be a source of measurement error, thereby, serving as a threat to validity.

In reality, the variability in students' appraisal of courses and teaching is contributed by distal (e.g., age, attitude, ethnicity, motivation, etc) and proximal (e.g., item difficulty, rater severity, the structure of rating scale, etc) facets. Distal facets are those variables that may have a mediated or indirect influence on the ratings provided by students (Eckes, 2015). Through a well-balanced and structured evaluation procedure, the negative effect of distal facets on the data can be minimised. Unlike distal facets, proximal facets have an immediate and direct effect on the rating scores awarded to the lecturers. Thus, the proximal facets, especially the person facet (i.e, lecturers in this study), plays a significant role in understanding the validity dynamics of the evaluation data (Eckes, 2015). Therefore, this study largely focuses on proximal facets and how they influence the variability in students' rating of courses and teaching.

In the context of students' appraisal of courses and teaching, which is a part of the rater-mediated assessment, the major proximal facets include instructor (in MFRM and not in GT), rater, and rating items (e.g., Börkan, 2017; Feistauer & Richter, 2016; VanLeeuwen, Dormody, & Seevers, 1999). Additional facet like occasion can be added if the data were obtained on a

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

number of occasions and thus, there is a vital reason to believe that occasion contributes to the variability in ratings (Li, Hou, Wang, Yang, Jian, & Wang, 2018; Quansah, 2020). In this study, the data to be used were obtained only on a single occasion and, hence, its omission as a facet. This study focused on the three (or two for GT) proximal facets since they are the basic ones that fit the evaluation data.

Non-response of students' evaluation of teaching quality survey

Students' appraisal of courses and teaching in the UCC is presently conducted through an online system. The aim, the researcher believes, is to reduce the time and resources used for carrying out such an activity. Literature, however, has found out that a high rate of non-response to online evaluation survey is a key challenge to the introduction of technology in students' appraisal of teaching and learning (Guder & Malliaris, 2013). McDaniel and Gates (2012, p.156) explained non-response bias as the "error that occurs as a result of a systematic change between persons who do and those who do not respond to a measurement instrument".

It is imperative to mention that the smallest non-response should be studied for several reasons. Non-response in students' evaluation of courses and teaching is increasing in recent times (Adams & Umbach, 2012). This issue of non-response has escalated as several tertiary institutions have transited from the traditional system of evaluation to the online mode of students' evaluation of teaching. This has been so, because of the need for institutions to reduce cost and improve data gathering and analysis. This transition has also led to a decreased rate of participation of students leading to a high rate of non-response (Dillman, Eltinge, Groves, & Little, 2002).

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

A unit of non-response produces an error that affects the quality of data obtained. In several instances, non-response in the survey is non-random (Dillman, Smyth, & Christian, 2009) suggesting that the non-response occurred for a reason. As students' non-response to an evaluation in schools increases, the probability that non-participants' views vary from participants' judgements also upsurges (Adams & Umbach, 2012). Thus, the accuracy of students' evaluation data may be questioned, if participants who failed to complete the evaluation systematically differ from those who completed the survey (McDaniel & Gates, 2012; Reisenwitz, 2015). That is, the validity of the responses in students' evaluation of teaching is affected and thus, leads to bias. The effects of non-response cannot be less emphasised because results from the survey seek to be generalised to the population and consequently, a reflection of the views of all who were eligible to participate in the survey (Groves, Fowler, Couper, Lepkowski, Singer, & Tourangeau, 2009).

The utilisation of data from students' evaluation of teaching, indeed, can be flawed by non-response. The potential for error in any form of survey increases as non-response upsurges. Thus, the worth of data and its use happens to reduce with declining rates of response (Groves et al., 2009). Data of this sort may have numerous uses at many institutions, therefore the quality of the data matters. Evaluation data may be used for promotion, reappointment, and instructional decisions (Porter & Whitcomb 2005). In this era of making decisions based on data, it is only vital to gather data that reflect the views of the larger population.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Generalizability theory and many-facet Rasch modelling

Generalizability theory (GT) offers a statistical theory and conceptual framework for assessing the dependability of a set of observed scores for a specified degree in diverse universes (Cronbach, Rajaratnam, & Gleser, 1963; Shavelson, Webb, & Rowley, 1989). In this study, the ratings of the students, regarding the quality of course content and teaching, are observed scores. By extension, these scores are only a sample of scores that the lecturers would have obtained under several conditions which serve the same purpose of estimating teaching ability. Generalizability Theory (GT) has its origins in Classical Measurement Theory (CMT) (Brennan, 2001b; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). In CMT, all errors of measurement are presumed to be random. Stated in variance terms, CMT divides the observed score variance into two fragments: (a) systematic or true variance and (b) random or error variance.

Generalizability Theory (GT), in contrast to CMT, discards the idea of a single undistinguishable error of measurement and somewhat postulates that the error of measurement occurs from multiple sources (Brennan, 2011). Specifically, GT estimates the amount of each error source distinctly and offers an approach for improving the dependability of behavioural measurements. One source of measurement error is a random error; other sources known as facets, such as raters (i.e. students in this study) and rating items, contribute to construct-irrelevant variances (Eckes, 2015). Additionally, GT evaluates the interaction between a number of facets and how this contributes to variability in observed measurements. For example, how students in a particular class systematically rate lecturers' teaching on a particular item was investigated. In

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

this case, the assumption was that there is an interaction between raters and items which was investigated using GT.

The Many-Facet Rasch Modelling (MFRM), on the other hand, is an extended form of the Rasch model (Rasch, 1980) which was advanced by Linacre (1989). For the basic Rasch modelling, test items and persons are assessed and positioned on an equal-interval scale, taking into consideration the differences in item difficulty and abilities of persons. Also, the results from Rasch modelling is not sample dependent regardless of the item used and ability distribution (Rasch, 1980). In practice, the MFRM is also a statistical approach used to investigate the variations in observable measurement. Particularly, MFRM is employed to evaluate characteristics of individual raters and how these raters influence the process of rating (McNamara & Knoch, 2012). MFRM evaluates the extent of influence of other sources of non-random errors like unreliable raters, inconsistency in ratings across occasions, and inconsistencies in the comparative difficulty of the items (Linacre, 1994).

Statistically, MFRM provides adequate knowledge on the extent to which the performance of persons, tasks, or raters is consistent with the expected predicted values from the model which has been generated in the analysis procedure (Linacre, 1994). The MFRM further permits "bias" analysis or circumstances of a possible interaction among different facets. For instance, the MFRM can provide information on whether a single rater tends to systematically score some category of individuals differently than the others, or some particular group of individuals performed systematically different on a specific item than they did on others (Linacre, 2003). In this study, for example, bias analysis can be performed to find out whether student raters consistently

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

rated some specific group of lecturers (e.g., lecturers who facilitate core courses) differently from other lecturers. This is an exploration of differential facet functioning, also known as differential item functioning when the facet in question is the item facet. This interaction (bias) analysis in MFRM can also examine differential rater functioning, differential person functioning, among others (Linacre, 2003).

Generalizability Theory (GT) and MFRM have both been found in the literature to be more appropriate in terms of assessing the dependability of observed data by estimating the true score and error score components of observed data, especially with rater-mediated assessment or evaluation (Brennan, 2011; Linacre, 2003). The utilisation of these two modelling approaches permits the identification of the various sources of measurement error associated with the observed data. GT and MFRM, again, compute for reliability indices which provides an idea of the dependability of the observed data (Brennan, 2011; Linacre, 2003).

Both GT and MFRM were included in this study because they complement each other despite the prominently different methodology both use in addressing the problem of measurement error (Kim & Wilson, 2009; Sudweeks, Reeve, & Bradshaw, 2005). It is understood that while GT sees the data principally from a group-level viewpoint, separating the sources of measurement error and calculating their extent, an MFRM analysis largely concentrates on individual-level information and, therefore, encourages functional examination into the functioning, or behaviour, of every individual component of the facets being considered (Linacre, 2001). For instance, GT can provide information regarding the similarity between the observed ratings of the

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

teaching quality of a group of instructors and the raw rating scores that identical groups of instructors may receive under similar conditions. MFRM rather will provide information on the extent to which each instructor was fairly rated. In this study, both group-and individual-level analyses were conducted and, thus, the need for applying both approaches.

Context of the study

In UCC, an office (a directorate, to be precise) has been created which is mandated with issues of quality assurance and have a responsibility of conducting students' evaluation of teaching and learning activities in the university. Formally known as the Academic Quality Assurance Unit (AQAU) from 2001 to 2006, the unit was mandated to provide prescribed institutional procedures and framework for quality provisioning in its academic programmes. The description of the focus and role of the AQAU was initially limited to only direct academic services (UCC, 2012). Currently called Directorate of Academic Planning and Quality Assurance (DAPQA-UCC) from 2006 to date, the responsibility of the Directorate has been extended to encompass all products and services that the university provides, which directly or indirectly affect her clientele (UCC, 2012).

Two of the major roles of DAPQA-UCC are to (a) intermittently assess research and teaching activities of departments and faculties, and (b) Monitor all resources and processes that indirectly or directly affect the teaching, research and learning environment (UCC, 2012). A major means by which the DAPQA-UCC performed this role is by permitting students to evaluate teaching and learning activities. Initially, this exercise was conducted through some

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

traditional means by moving from one lecture hall to another and administering to students the evaluation forms to appraise the quality of courses and teaching.

In recent times, however, students' evaluation of courses and teaching is being conducted online through students' portal accounts. The evaluation exercise is conducted every semester for respective courses and this is usually carried out close to the end of the semester where students have been taken through some reasonable period of instruction. This data on students' appraisal of courses and teaching are used for lecturers' professional development, course content modification, promotion and confirmation decisions, among others (Gyimah, Kwarteng, Anane, & Nkrumah, 2016).

This study attempts to understand whether evaluation data obtained from students regarding the quality of teaching and course contents in UCC is something that can be relied on for decision making. This is done by investigating how basic proximal facets (factors), as per the data, influence the quality of data obtained from students. Thanks to the measurement scholars, GT and MFRM provide a framework for the management of UCC to know how valid evaluation data is. CMT was excluded because of the flaws associated with its use; these flaws are addressed by GT. Indeed, there is a need for the conduct of this study to investigate the validity of students' appraisal of courses and teaching using GT and MFRM in UCC.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Statement of the Problem

The issue of investigation aroused from a number of perspectives. Thus, the problem is stated based on the issue, context, and theory and methodological gaps.

The issue gap

In UCC, data from students' evaluation of the quality of the course and teaching serve both formative decisions (i.e., lecturer development decision, and improvement and/or modifications in teaching and course contents) (UCC, 2012) and summative decisions (i.e. lecturers' appointment and promotion) (UCC, 2015). While little evidence has been established to argue out that a lecturer may be denied 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 these evaluation data, the quality of such data needs to be paramount and this becomes a concern not only for university administrators but also for the instructors and researchers.

In a cross-sectional survey among lecturers of all ranks in UCC, the lecturers lamented that students poorly rated the quality of teaching of instructors who were strict on them (Gyimah et al., 2016). Although the recognition and acceptability of students' appraisal data were high among lecturers, some lecturers argued that the students do not have a good value-judgment in assessing their teaching activities and may also not be responsible enough to take up this 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 was supported by another study conducted in UCC by Kwarteng, Doku, Matta, and

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Doh-fia (2014), it was reported that there were fluctuations in lecturers' overall ratings provided by the students over 5 years (i.e., from 2008/2009 academic year to 2012/2013).

In 2015, students' evaluation of courses and teaching in UCC moved from the traditional system of appraisal (face-to-face data collection) to online mode (through students' portal accounts) (UCC, 2012). My experience as a student in UCC for the past 7 years has revealed to me that some proportions of the students rarely respond to the evaluation of courses and teaching online survey organised by the university. The researcher has encountered a situation where students feel that instructors may penalise them when they genuinely evaluate the instructor. These students end up not responding to the evaluation form at all or provide an inaccurate assessment of the instructor. Both the latter or former has validity implication for the data provided.

Corroborating my experiences, several and growing pieces of evidence have found increasing non-response among students in the appraisal of teaching and courses in universities (Adams & Umbach, 2011; Dillman et al., 2002; Porter & Whitcomb, 2005; Groves et al., 2009). Non-response in an online survey, especially in the case of students' appraisal of teaching and courses, produces errors that affect the validity of students' evaluation data (Dillman et al., 2009; McDaniel & Gates, 2012; Groves et al., 2009). With this evidence of the grave consequences of non-response, there is a need to investigate non-response in students' appraisal and how these non-responses (if any) have any influence on the quality of the evaluation data provided by students. In this study, the influence of non-response on the validity of the online evaluation survey was investigated.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Contextual gap

Research works on the sources of variability and the validity of students' evaluation of teaching and courses in schools have been extensively conducted across America, Europe, and Asia. Taking America, for example, such studies have been carried out in New Mexico over 20 years ago (VanLeeuwen et al., 1999), South-eastern United States (Reisenwitz, 2015), North Carolina (Adams & Umbach, 2012) and in other parts of the US (Zhang, Fike, & DeJesus, 2015). In Europe, similar studies have been conducted in Germany (Feistauer & Richter, 2016), Italy (Braga, Paccagnella, & Pellizzari, 2011), Denmark (Sliusarenko, 2013) and other regions of Europe (Goos & Salomons, 2017). On the Asian continent, studies were reported in Malaysia (Fah & Osman, 2011; Samian & Noor, 2012; Sok-Foon, Sze-Yin, & Yin-Fah, 2012), Turkey (Börkan, 2017; Üstünlüoğlu & Can, 2012), Pakistan (Raza & Irfan, 2018), and China (Li et al., 2018).

It is only recently that two studies on the sources of variability and validity of students' appraisal of courses and teaching were conducted in Africa; specifically, in Ghana (Quansah, 2020) and South Africa (Ogbonnaya, 2019). The study in South Africa was, however, conducted at the secondary school level, indicating relatively low investigations on the subject matter in higher education in Africa. In Ghana, it has been found that the sources of variability, as well as, the reliability and validity of students' evaluation is still at the infant stage (Quansah, 2020). While much is unknown on the subject matter in Ghana, the dynamics in the accuracy of students' appraisal data might change from one country to another or even from one institution to another in the same country. This is because these students find themselves in different socio-cultural

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

environments and thus, variables in such milieu influence who is appraising and how the evaluation is carried out (Eckes, 2015). This explains the inconsistencies in the results of previous literature; whereas, some studies found inaccurate results on students' appraisal of course and teaching (e.g., Börkan, 2017; Feistauer & Richter, 2016), others found that students provided valid evaluations (e.g., Samian & Noor, 2012; VanLeeuwen et al., 1999).

In effect, not much is known in Ghana, and UCC to be precise, regarding the validity of students' appraisal of courses and teaching. This study is conducted in UCC for two reasons: (1) a number of studies have been conducted in an attempt to understand the dynamics in student ratings of instructors and their courses (Gyimah et al., 2016; Kwarteng, Doku, Matta, & Doh-fia, 2014) and (2) the researcher has experienced the evaluation system in the study setting which serve as a motivation to conduct such a study. Although, the investigator is a student within the institution, the statistical procedures employed are so objective that the investigator has no influence on the results.

Theory and methodological gap

A review of existing literature on issues of validity and dependability of students' appraisal of courses and teaching around the globe have shown the utilisation of three predominant approaches: CMT, GT and MFRM. The majority of these studies employed CMT which is the foundation of most, if not all, measurement theories (e.g., Adams & Umbach, 2012; Braga et al., 2011; Fah & Osman, 2011; Goos & Salomons, 2017; Ogbonnaya, 2019; Samian & Noor, 2012; Sliusarenko, 2013; Raza & Irfan, 2018; Reisenwitz, 2015; Sok-Foon et al., 2012; Üstünlüoğlu & Can, 2012; Zhang et al., 2015). It is in very few instances that studies explored the validity and reliability of students'

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

environments and thus, variables in such milieu influence who is appraising and how the evaluation is carried out (Eckes, 2015). This explains the inconsistencies in the results of previous literature; whereas, some studies found inaccurate results on students' appraisal of course and teaching (e.g., Börkan, 2017; Feistauer & Richter, 2016), others found that students provided valid evaluations (e.g., Samian & Noor, 2012; VanLeeuwen et al., 1999).

In effect, not much is known in Ghana, and UCC to be precise, regarding the validity of students' appraisal of courses and teaching. This study is conducted in UCC for two reasons: (1) a number of studies have been conducted in an attempt to understand the dynamics in student ratings of instructors and their courses (Gyimah et al., 2016; Kwarteng, Doku, Matta, & Doh-fia, 2014) and (2) the researcher has experienced the evaluation system in the study setting which serve as a motivation to conduct such a study. Although, the investigator is a student within the institution, the statistical procedures employed are so objective that the investigator has no influence on the results.

Theory and methodological gap

A review of existing literature on issues of validity and dependability of students' appraisal of courses and teaching around the globe have shown the utilisation of three predominant approaches: CMT, GT and MFRM. The majority of these studies employed CMT which is the foundation of most, if not all, measurement theories (e.g., Adams & Umbach, 2012; Braga et al., 2011; Fah & Osman, 2011; Goos & Salomons, 2017; Ogbonnaya, 2019; Samian & Noor, 2012; Sliusarenko, 2013; Raza & Irfan, 2018; Reisenwitz, 2015; Sok-Foon et al., 2012; Üstünlüoğlu & Can, 2012; Zhang et al., 2015). It is in very few instances that studies explored the validity and reliability of students'

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

evaluation of teaching using GT (e.g., Feistauer & 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).

Over the years, the CMT approach to the study of variances in and validity of psychological measurement has been criticised (already discussed earlier in this write-up) and as a result, the entire CMT framework was expanded into what is known as GT (Brennan, 2001a). Although CMT and GT are similar in some respect, GT has been found to provide richer, comprehensive and more accurate information in the validity of psychological measurements, especially, in situations of rater-mediated evaluations and assessment (e.g., students evaluating courses and teaching or a teacher grading an essay test) (Brennan, 2011).

From the foregoing, the results of studies that employed CMT are seen as inaccurate and narrow regarding their contributions to literature in the area. Cronbach et al. (1972) and Brennan (2001b) gave some reasons for this: (1) GT allow the disentangling of the single error component in CMT into multiple sources of error, (2) GT has different decision study designs and allows the distinction between random and fixed measurement facets, (3) GT permits a multivariate framework which extends reliability to multiple universes of generalisation, and (4) GT bridges the gap between reliability and validity (Kane, 1982). The aforementioned discourse does not suggest that GT should completely replace CMT, but in the case of rater-mediated evaluation which is the focus of this study, GT is more appropriate and preferable (Brennan, 2011).

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Similarly, studies conducted using GT, as compared to 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 group-specific analysis of the variances influencing the scores for all object of measurement (e.g., lecturers, as in this study) (Linacre, 1994). Unlike GT which has no implication for an individual object of measurement except the number of observations made, MFRM focuses on an individual object of measurement and as well further provides quality control statistics for each element on a common and equal-interval scale (Linacre, 1994). While GT, for example, can tell how different lecturers were rated by 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). While MFRM can provide a fair score distribution for each object of measurement, GT cannot (Linacre, 1994). Despite these discrepancies, scholars have shown that both approaches are unique, relevant, and complementary in their utilisation to investigate the dependability of behavioural measurement (Lee & Cha, 2016; Linacre, 1995; Sudweeks et al., 2005).

Researchers in the measurement field have extensively compared the utilisation of GT and MFRM with regards to key research questions, modelling, design, data collection approaches, limitations and statistical results (e.g., Brennan, 2001b; Kim & Wilson, 2009; Linacre, 1993, 1995, 2001). These researchers, after comprehensive simulations exercises, suggested that GT should be first conducted to get an understanding of the data and then subsequently, conduct MFRM to obtain more detailed results. For instance, GT can identify whether raters were consistent or not but cannot tell whether raters

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

were more severe or lenient; this information can be provided in the MFRM analysis.

Merging both GT and MFRM also supports recent developments in validity theory which recommends multiple sources of validity evidence to be gathered, assessed, and combined into validity argument to support score interpretation and utilisation (Fan & Bond, 2006; Kane, 2012). Given the benefits of combining GT and MFRM in assessing the validity of students' appraisal of courses and teaching, available literature has revealed that there appears to be no study conducted in the area which employed both approaches. This question the comprehensiveness of the results presented by previous studies; one approach (either MFRM or GT) cannot provide all information about the data. With this, it is possible to mention that issues surrounding the validity of students' appraisal data are not well understood. In this study, the researcher examined the validity of students' evaluation of courses and teaching using GT and MFRM.

Purpose of the Study

The main purpose of this study was to examine the validity of students' evaluation of teaching and courses in UCC through the lenses of GT and MFRM. In terms of the specific objectives, this study utilised GT to:

1. examine the sources of measurement errors (rater and items) in students' appraisal of courses and teaching;
2. establish the extent of dependability of students' ratings of the quality of teaching and course contents;
3. find out how many raters and items are needed to obtain an optimum level of dependability in students' appraisal of courses and teaching;

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

4. assess how non-response influence the validity of students' evaluation of courses and teaching;

The study also employed MFRM to evaluate:

5. the behaviours of students (e.g., halo effect and severity effect) in the appraisal of courses and teaching;
6. how effective evaluation items are in the measurement of the quality of course content and teaching;
7. whether the lecturers are fairly rated in terms of their quality of teaching;
8. whether there are biases in students' appraisal of courses and teaching.

Research Questions

The study was guided by a number of research questions. With regards to the use of GT, the following questions were asked:

1. What are the sources of measurement errors (e.g., raters and items) in students' appraisal of courses and teaching?
2. What is the extent of dependability of students' ratings of the quality of teaching and course contents?
3. How many raters and items are needed to obtain an optimum level of dependability in students' appraisal of courses and teaching?
4. How does non-response influence the validity of students' evaluation of courses and teaching?

Based on the utilisation of MFRM, the following questions were asked:

5. What are the behaviours of students (halo effect, and severity effect) in the appraisal of courses and teaching?
6. How effective are evaluation items in the measurement of the quality of course content and teaching?

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

7. How fair are lecturers rated in terms of the quality of lecturers' teaching?
8. What are the biases in students' appraisal of courses and lecturers' teaching?

Significance of the Study

The findings of this study contribute to the literature in the field of educational quality and quality assurance in higher education institutions. The study extends the knowledge provided by previous studies concerning students' appraisal of the quality of courses content and teaching. For instance, previous studies largely adopted approaches (to the study of the validity of students' evaluation data) that did not offer a comprehensive understanding of such data. It must be mentioned that those studies provided some basic understanding of students' appraisal data, only that little information was provided to understanding the dynamics in students' ratings of lecturers. This study, therefore, combine two approaches to explore the validity of students' evaluation data to improve the depth of knowledge in such areas and thereby serve as a foundation for other areas of study to emerge.

Furthermore, the study offers the utmost benefit to the management of UCC in many ways. First, the outcome of this study provides the university with an opportunity to understand the dependability of data from students' evaluation of courses and teaching. With this knowledge, the management of the university can make an informed decision on how to use such data (i.e., whether to use it for formative decisions or summative decisions). Further, the study offers DAPQA-UCC information on the sources of measurement error regarding the credibility of data on students' ratings of courses and teaching. This information will help the Directorate to work towards reducing the errors identified to

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

influence the dependability of the data. For example, a new instrument can be developed for the students' appraisal exercise in the institutions since the study found out that the existing instrument (or items) was relatively less effective for the purpose it serves.

This study is relevant to the Academic Board of UCC, DAPQA-UCC, and the university at large in terms of introducing to them the two statistical measurement theories (GT and MFRM) which can be adopted for assessing the quality of evaluation data in the future. This is vital because the MFRM, for instance, can provide fair scores for all the lecturers and these scores can be used for decision making instead of the original scores provided by the students. If this becomes feasible and effective, the statistical theories can be applied to other operations of the university such as examining how reliable internal and external examiners score thesis/dissertations of students at the postgraduate level.

This study is relevant to other educational institutions that use students as raters for the evaluation of courses and teaching. The outcome of this study serves as a prompt for other institutions of higher education to pay attention to the quality of data on quality assurance. Thus, this study offers well-balanced detailed and scholarly information on how appraisal data from the student should be understood and utilised in higher education institutions, not only in Ghana but also in African educational institutions and even beyond. This will ensure that better quality assurance decisions are made in higher educations if only administrators of these institutions understand the nature of data provided by important stakeholders like students.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Delimitations

This study was limited to UCC (all colleges except for College of Distance Education) for two major reasons: (1) investigator have personal experience about the subject matter under investigation and this became the starting point of this research, (2) it is the only institution in Ghana where there is empirical evidence which points to the fact that lecturers have little trust for the data on students' evaluation of courses and teaching.

This research examined the dependability of students' evaluation of the quality of course content and teaching by examining the following sources of measurement error: evaluation items, lecturers (for MFRM only), rater behaviours, and non-response. Although there are a number of measurement theories, this study primarily focused on two: GT and MFRM which were chosen based on their efficacy. The study made use of secondary data from the university which were provided by the students for administrative purposes.

The study did not explore how certain key variables interact, such as gender, the programme of study, and level of study, interacts with the rater/person/item facets to bias in students' appraisal of courses and teaching. Also, the procedure utilised for the evaluation of non-response in students' appraisal of courses and teaching only established whether there is an association between students' non-response, measurement error and validity of responses.

Limitations

Despite the significance of the study, there were some limitations. First, the data used were only for a single semester (i.e., the second semester of the 2019/2020 academic year) and thus, the results may not be representative

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

enough for generalisation. It must be stated that the data had a high rate of students' non-response and this suggested that the majority of the students failed to respond to the evaluation survey. Therefore, the analysis only covered about half of the students and this may not reflect the student population. This may limit the generalisation of the research.

Further, there are some variables which were not controllable by the researcher. Factors such as sample size, gender distribution, lack of demographic variables for the data, replication of responses, and biases in course representation are key limitations to this research. Also, the data was dominated by undergraduate students and this might affect the generalization of the results. The fact that the data was collected for a purpose other than this research might affect the generalization of the results.

The outcome of this research may be limited in terms of its applicability to other institutions of higher education. This is because of the diversities in the evaluation of teaching in different educational institutions. That is, different educational institutions may have different evaluation instrument and items and may even attempt to measure traits of teaching and activities which are not taken into consideration in UCC. This makes it difficult to apply the findings of this research to other higher education enterprises. It must be acknowledged that irrespective of this limitation, this research may serve as an "eye-opener" for other tertiary institutions in Ghana and beyond.

Definition of Terms

Validity: In the context of this research, validity is the extent to which scores provided by students regarding course/lecturer appraisal reflect the quality of teaching of a lecturer and/or quality of course content.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Facet: A facet is a set of related conditions of measurement. With regards to this study, facets include those proximal factors which can introduce errors in students' ratings of the quality of teaching and course content. The facets used in this research were items, raters (students), and lecturers (considered as a facet in the framework of MFRM and not GT).

Halo effect: Halo effect, as used in this research, signifies the propensity of a student to offer equivalent ratings for a lecturer/course on practically different criteria. Fo example, a lecturer's class attendance behaviour may influence a student to rate the lecturer excellently or poorly on the lecturer's quality of teaching.

Item difficulty: Item difficulty, in this research, is used in the item response theory sense, particularly the polytomous response type. In the context of this study, more difficult items are items on which lecturers have a higher probability to obtain a low score, irrespective of their ability. Less difficult items are those items on which lecturers have a higher chance of receiving a high score despite the lecturer's ability.

Rater severity: The degree to which a student assign a very high or low rating for a lecturer or course than those scores provided by other students evaluating the same lecturer or course.

Organisation of the Rest of the Study

The entire research report is catalogued into five chapters. Following this chapter, which is the introductory chapter, is the literature review (chapter two) where pertinent and related literature is extensively discussed. The review highlighted the explanation of key concepts in the area of students' evaluation of teaching quality; theoretical framework such as GT and MFRM; theoretical

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

review such as Kirkpatrick's evaluation model; review of previous studies; and conceptual framework. Chapter three discussed the research methods which includes the research paradigm, research approach, GT and MFRM designs, study area, population, sampling procedure, data collection instrument, data collection procedure, ethical considerations and data management, and data processing and analysis procedure. Chapter four presents the analysis of the data extracted as well as the discussion. This thesis report concluded with a summary of the study (including the findings), conclusions, recommendations for policy and practice, and suggestions for further studies.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

CHAPTER TWO

LITERATURE REVIEW

Following the introductory chapter of this study, the literature chapter is next. This chapter discussed significant literature which was related to the issue of students' appraisal of courses and teaching. These issues included theoretical reviews and frameworks, a brief description of key concepts, empirical reviews and a conceptual framework that was later designed to guide the conduct of the research.

Brief Description of Literature Search Process

This review adopted a selective approach that focused on articles and journals of particular relevance to the work. Various measurement sources of books and articles were directly targeted. The search processes were conducted largely via electronic database including Scopus, Google Scholar, Web of Science (WoS), general web searches, measurement journals search, and bibliographies. Searches included terms and phrases such as 'students' appraisal of teaching', 'students' appraisal of teaching in higher education', 'students' evaluation of teaching', 'generalizability theory', 'many-facet Rasch measurement', 'validity of students' evaluation of teaching', 'dependability of students' appraisal of instruction', reliability of students' evaluation of teaching', 'generalizability theory and students' evaluation of teaching', 'many-facet Rasch modelling and students' evaluation of teaching'.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

First, hundreds of information from secondary sources were identified. Then, the references were carefully examined to 'fish out' for the primary literature to be used. That is, citations and references were reviewed for potential articles that might be missed. The information which was not directly related to the issue under investigation were eliminated. For instance, the literature on the use of MFRM in analysing open-ended statistics questions by Güler (2014) was archived through the search process but was excluded from the review because the study was unrelated to this study. Results were refined to the current topic of interest, subject area and titles.

Based on the search and synthesis of the gathered literature, the literature review was categorised under four major sub-sections: a review of key concepts, theoretical framework, conceptual framework, and empirical review.

I. Review of Key Concepts

1. Concept of students' appraisal of teaching
2. Dimensions of students' evaluation of teaching
3. Uses of students appraisal of teaching data

II. Theoretical Framework

1. Generalizability theory
2. Many-facet Rasch modelling

III. Theoretical Review

1. Validity Theory
2. Kirkpatrick evaluation model
3. Selection Bias Model

IV. Review of Related Studies (Empirical Review)

1. Previous Studies which employed CMT

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

2. Empirical studies which employed GT
3. Existing studies which employed MFRM

V. Conceptual Framework

Review of Key Concepts

Concept of students' appraisal of teaching

With respect to students' university experiences, examining whether an instructor or teaching process is bad or good can be likened to the measurement of quality of service in the business world (Curran & Rosen, 2006). This is particularly the situation with regard to students' appraisal of teaching in educational institutions. The concepts behind students' appraisal of teaching have been extensively discussed since the early 1920s (Otani, Kim, & Cho, 2012), and explicitly, several pieces of research have been conducted on issues of students' evaluation of learning experience, teaching process, instructional quality and quality assurance related concepts (Marsh, 2007).

Students' appraisal of teaching is a fairly new concept which was introduced and utilised interchangeably with numerous previous expressions such as students' evaluation of teaching effectiveness (Marsh, 2007), student appraisal of instructor performance (Chuah & Hill, 2004), students' evaluation of educational quality (SEEQ) (Coffey & Gibbs, 2001; Grammatikopoulos, Linardakis, Gregoriadis, & Oikonomidis, 2015; Lidice & Saglam, 2013), student course satisfaction (Betoret, 2007; Bolliger, 2004; Rivera & Rice, 2002), students' evaluation of instruction (Clayson, Frost, & Sheffet, 2006), or just student course evaluation (Anderson, Cain, & Bird, 2005; Babcock, 2010; Bembenutty, 2009; Chen, 2016; Duggan & Carlson-Bancroft, 2016).

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Notwithstanding the disparities in the concepts, they have a common underlying objective.

Based on the definition and classification of fundamental higher education concepts outlined by the United Nations Educational, Scientific and Cultural Organization (UNESCO), students' appraisal of teaching is explained as:

> "... the process of using student inputs regarding the general activity and attitude of instructors. These observations permit the overall assessors to examine the extent of conformability between the expectations of the student and the actual teaching methods of instructors. Student appraisals are expected to provide insights concerning the instructor's attitude in a class and/or the abilities of an instructor ..." (UNESCO, as cited in Vlăsceanu, Grünberg, & Pârlea, 2004, pp. 59-60).

This explanation highlights three key characteristics, namely, the appraisal of the instructor (the instructor), the instructional process (general activity and instructional methods), and the instructional outcomes from the lenses of the students (expectations of the student). This reflects the framework for appraising service marketing, where the instructor is linked to the "search" features, the instructional process to the "experience" features, and the instructional outcomes to the "credence" characteristics (these dimensions have been discussed in the next section).

The concept of students' appraisal of teaching, as used in this research, is seen as the process by which students assess the teaching activities and course contents relative to their learning experiences. It is important to emphasize that

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

these features that students evaluate include the instructors' ability to explain issues to students, guide their learning, and drive class discussions. The instructors' competencies in unpacking course contents and making courses relevant to students are also key in the assessment process. In UCC, for example, the evaluation extends evaluation feature from the mode of delivery and course contents to instructors' punctuality to lectures and assessments given by the lecturer.

Dimensions of students' evaluation of teaching

The dimensions of students' appraisal of teaching are those key elements that form the entire evaluation system in the context of teaching. Prominent in literature is the following three dimensions: teacher, teaching process and learning outcomes. These dimensions have been discussed in subsequent sections of this report.

The teacher

One of the first aspects of students' appraisal of teaching emphasises the teacher, or perhaps the perception of students regarding the characteristics of the teacher (Hu, Hung, & Ching, 2015). It is worth stating that the emphasis on the instructor/teacher, most of the time, results in biases on the part of the students during the appraisal process. For instance, instructors' physical appearance can influence their evaluations scores (Bonds-Raacke & Raacke, 2007; Hultman & Oghazi, 2008). Supporting this assertion, Felton, Mitchell, and Stinson (2004), in their research on university lecturers' evaluation by students, showed that physically attractive lecturers received high evaluation ratings. Other studies like that of Sohr-Preston, Boswell, McCaleb, and Robertson (2016) found that instructor's physical looks, age, and gender are key

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

when students are rating their instructors. Related to this study, Boring, Ottoboni, and Stark (2016) discovered that younger male instructors were found to receive higher scores in their evaluation as compared to older lecturers, and this was influenced by physical appearance.

As a way of clarification, teacher qualities have been explained in the literature as straightforwardly observable and the most common physical characteristics an instructor may have (Curran & Rosen, 2006). Furthermore, these search features are outward and can be assessed in relation to comparable options (Lubienski, 2007). The key issue is the ability of the students to accurately evaluate these qualities being searched for. This suggests that there may some expected criteria against which the students conveniently evaluate the instructors (Clayson, 1999).

Also, literature in psychology has long established the effect of attractiveness on the judgment of individuals (Brownell & Swaner, 2009). This indicates that evaluators may equate physical appearance to the degree of the constructs being measured. This concept is usually referred to as the halo effect. This halo effect describes why instructors' attractiveness has a tendency to influence their ratings regarding their quality of instruction (Felton et al., 2004). Additionally, students may also have in mind some preference regarding the physical situation of the classroom (Douglas & Gifford, 2001; Hill & Epps, 2010), which consciously or unconsciously affects their general satisfaction level.

Teaching process

Aside the teacher, the second component of students' appraisal of the teaching system is the teaching process. In practice, the majority of students'

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

appraisal exercise conducted in various institutions target measuring this construct (Algozzine et al., 2004; Wachtel, 2006). The objective of students' evaluation of teaching is to assess the instructor's effectiveness in teaching (Marsh, 2007). This approach has been employed throughout educational institutions for decades, but its usefulness, validity, and reliability are still being tested (Arreola, 2007; Marsh, 2007). This is because of the complexities associated with the teaching activity and as such the multidimensional view of the teaching process (Marsh, 2007). However, it has been demonstrated that students' evaluation of teaching significantly contributes to instructor growth and development and thereby, improving the teaching and learning experiences of students (Marsh, 2007). This formative function reinforces the need for the evaluation of some or all aspects of teaching so that improvement can be guaranteed in the overall quality of the teaching process (Berk, 2005).

Evaluation of the teaching process, as have been earlier indicated, is a complicated and complex undertaking, which need complete insight into how students came to conclusions concerning the quality of course content and teaching (Curran & Rosen, 2006). Technically, being registered for a university course in a full-time mode would mean that the student would attend a class every week, which reflects recurrent service that is important for later judgement of the teaching situation (Lubienski, 2007).

These features of the services received are not readily observable and can only be evaluated when the complete service experiences have ended (normally close to the end of the course) (Curran & Rosen, 2006). Nevertheless, since such experiences are recurrent, it becomes difficult to know whether the student appraisals are resulting from the complete experience of the course or

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

just one single situation that made a long-lasting impression (Curran & Rosen, 2006). Moreover, students go to class with their colleagues, so other students are participating in the service delivery simultaneously. Hence, the interactions of these students within the class might either improve or reduce the teaching quality measurement (Fisk, Grove, & John, 2014).

Pietersen (2014) proposed having a psychological contract between the students and the teachers, to offer unambiguous study objectives and eliminate false expectations. Furthermore, evaluations can be carried out in the middle of a particular semester to give the instructor a sufficient chance in addressing students' doubts and to re-structure the shared contract depending on the abilities of the students (Brownell & Swaner, 2009). Additionally, as the objective of students' evaluation of teaching is to offer formative recommendations for instructors to enhance teaching effectiveness, it is also practical to include declarations on the delivery of formative instructions and on how course designs influence student learning experiences (Kuh, O'Donnell, & Reed, 2013).

In the nut shell, it has been outlined in the quality assurance literature that appraising teaching effectiveness or the teaching process is a multidimensional task. It is multifaceted and typically reflects students' experience qualities. Due to the several biases linked to students' evaluation of teaching, the students must become aware of the expectations of the teaching process in order to accurately make a judgement about this process. This notwithstanding, it appears to be difficult to satisfy all customers (students) due to individual expectations and perceptions; but one thing needs to be clear, and

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

that is, students have the ability to provide similar ratings when the evaluation process is structured very well (Quansah, 2020).

Learning outcomes

The final aspect of the students' appraisal of the teaching system is the learning/instructional outcomes, more explicitly, the achievement of objectives. It has widely recognised that goals are significant contributors to academically applicable learning outcomes, while also concentrating on the motivational components motivated by mastery and performance-approach goals (Harackiewicz, Barron, Pintrich, Elliot, & Thrash, 2002). Simply put, if students undoubtedly understand the skills important for future employment, while also understanding that enrolling in a particular course will help them to have mastery of those skills, they will pay much attention to such a course. It has been shown that students are more involved with their academic work when future professional consequences are known (Greene, Miller, Crowson, Duke, & Akey, 2004). Nevertheless, many students in reality are indeterminate in their educational goals and are in danger of dropping out of school (Mäkinen, Olkinuora, & Lonka, 2004).

The kind of education provided by the universities is characterised by high credibility properties (Curran & Rosen, 2006). Credence qualities are those features that are not obvious, can be never be completely appreciated or known by students (Lubienski, 2007), and might, thus, be not possible to assess (Curran & Rosen, 2006). Credence characteristics are commonly found in goods and services that are characterised by high complexity levels like the teaching and learning processes. More significantly, even when the service has been used (in this instance, when a student have graduated from the university), the client

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

(student) may still have difficulties evaluating the service they received some time ago (Voss, Gruber, & Szmigin, 2007). In the institutions of higher education, the true worth of a course may be appreciated by a student only after the knowledge and skills attained are utilised in employment situations, which might be numerous months or even many years after the service has terminated (Curran & Rosen, 2006; Galetzka, Verhoeven, & Pruyn, 2006).

Uses of students' appraisal of teaching data

Research studies account for several positive benefits from the use of students' evaluations of teaching. One benefit is the development or self-improvement of instructors (Stapleton & Murkison, 2001; Simpson & Sigquaw, 2000). Faculty can take the formative results from teaching assessments and change their classroom performance or they can strengthen good practices (Lewis, 2001). Faculty will benefit from student evaluations if: (1) they learn new information from the results; (2) they appreciate the results; (3) they can change their classroom behaviours.

Most universities conduct student evaluations at the end of the semester (McNulty, Gruener, Chandrasekhar, Espiritu, Hoyt, & Esminger, 2010). Some researchers suggest that faculty can benefit even more if they are given the formative results from early semester evaluations (McGowen & Osguthorp, 2011). This allows the instructor to change classroom behaviours and improve teaching during the semester rather than waiting until the end. This will directly impact student learning. It also gives the students a chance to design their instruction. Assessments made early in the semester are more useful to faculty compared to a global assessment at the end of the semester (McNulty et al., 2010).

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Administrators can use the summative results to make decisions about hiring, promotions, pay raises, teaching awards, tenure, and professional development (Beran & Violato, 2005). Beran, Violato, and Kline (2007) discovered that university administrators used evaluation results for personnel decisions. Some higher education institutions use the written comments on faculty evaluations to nominate instructors for teaching awards and for future reference for promotion or tenure (Gravestock & Gregor-Greenleaf, 2008). Some schools publish the results and students can use the information to select courses (Beran, Violato, Kline, & Frideres, 2007).

Theoretical Framework

The theoretical framework of every study highlights the structure that supports and embraces the research and as well defines the conduct of the research. Two major theories served as the theoretical framework: Generalizability theory and Many-facet Rasch modelling.

Generalizability theory

Generalizability Theory (GT) is an arithmetic theory concerned with the reliability of behavioural measurements. The GT is an extension of the Classical Measurement Theory (CMT) (Cronbach et al., 1972). The CMT operates on the assumption that each test score or observation comprises a true score and an error score that generates a single dependability coefficient for a set of equivalent observations. This hypothesis may be practical when the parallel forms are equalised cautiously. This hypothesis, on the other hand, becomes unrealistic where variances or average scores are dissimilar or when there are heterogeneous items on the form.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

From the perspective of internal consistency, reliability appears to be low in a multidimensional measurement; nevertheless, parallel forms and test-retest reliability estimates may be high at the same time. Cases like this present contradictions and restrictions of the CMT model of reliability. The GT was designed as a less rigid approach that seeks to eliminate this restriction; and to help in the analysis of error which has the probability of emanating from potential sources of variability like the tasks, rater, items, measurement form, time. The GT framework puts the different sources of variation altogether and simultaneously compute an all-inclusive dependability reliability estimate. Furthermore, the GT removes the traditional variations between validity and reliability (Allal & Cardinet, 1997).

Shavelson and Webb (1991) believed that GT is a protracted form of the CMT due to four reasons: 1) The GT has the ability to estimate several sources of variability in a single computation; 2) The use of GT guarantees the estimation of the magnitude of each source of variability; 3) GT permits the computations of two different errors of measurement and reliability coefficients and thus, making it possible to take relative and absolute decisions; 4) GT allows for realistic measurement decisions to be made to reduce errors measurement to the barest minimum based on some specific purpose (D-study).

It is worthy to state that factors as time, tasks or items, and rater is known as the facets or source of variability in the framework of GT (Brennan, 2001). Stated differently, facet is a concept that is substituted for all sources of probable measurement errors. Therefore, it is preferable to reduce the degree of variation related to the source of variation as low as possible rather than maximising (Alharby, 2006). Each source of variability has levels that are referred to as the

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

conditions of measurement. For this study, for example, 13-items were used for the measurement of the teaching construct and thus, the item facet had 13 conditions of measurements. Similarly, the rater facet had over 2700 raters and so the conditions of measurement for the rater facet is over 2700. In general, the potential conditions of measurement for any random facet are deemed infinite in magnitude. The selected conditions of measurement which were admissible to the investigator are known as the universe of admissible observation.

Another concept that is worthy of explanation is the universe of generalization. This concept signifies the set of conditions the investigator would wish to generalise to. In simple terms, persons act as the focus of the measurement on which conclusions are based. Consequently, persons are not considered as a source of variability because variations contingent on persons is always preferred.

Universe score is another concept that needs to be explained. By definition, a universe score is a measurement score that reflects the average of scores attained from the universe of admissible observations for the sources of variation. The universe scores variance is analogous to the true score variance as used in the framework of CMT; hitherto as dissimilar from it, two distinct error variances are obtained in the GT framework. The difference originates from the notion that GT allows two decisions to be made. Both the absolute and the relative error variances are computed and interpreted in the GT. The relative error variance, in this case, is analogous to the error variance in the CMT framework (Shavelson & Webb, 1991).

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Framework of GT

The framework of GT is shown in Figure 1

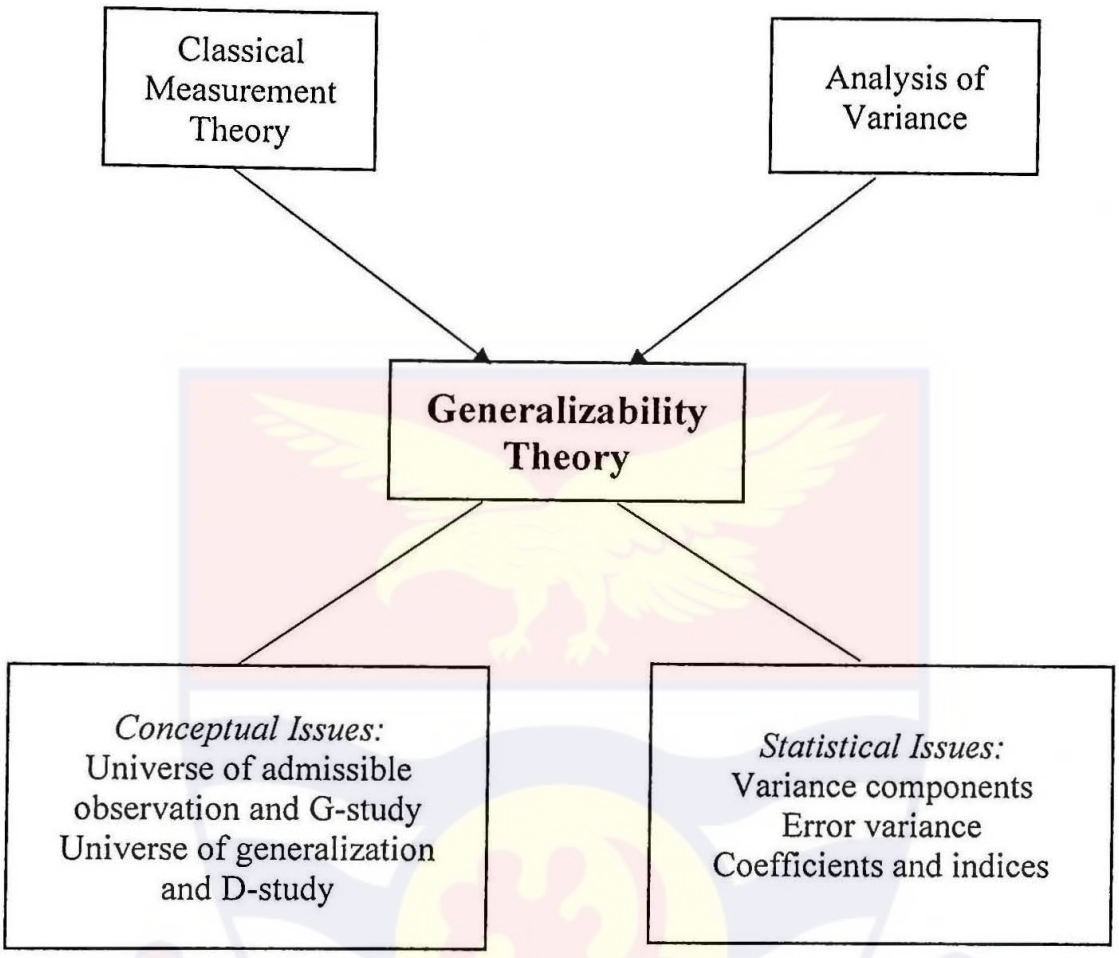


Figure 1- Parents and Conceptual Framework of GT

Even though CMT and analysis of variance (ANOVA) are seen as the parents of GT, the child is equally more and less than the simple combination of its parents, and understanding GT necessitates an insight of more than its ancestry (Figure 1). Though GT loosens CMT, not all characteristics of CMT are unified into the framework of GT explained by Feldt and Brennan (1989). Likewise, not every component of ANOVA is applicable to GT; undeniably, some viewpoints on ANOVA contradicts GT. Additionally, the ANOVA concerns emphasised in GT are dissimilar from those that prevail in several pieces of literature on ANOVA and experimental designs. To be more specific, GT focuses on variance components and their computations.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

G-study

In GT, a coefficient that is referred to as the generalizability coefficient is computed. This coefficient does not re-interpret the notion of reliability though analogous to the coefficient in CMT. The GT framework bridges the gap between reliability and validity. In GT a universe, its conditions of observations and variability sources are explained from the structure described in the original field of the concept of validity (Shavelson & Webb, 1991). The aim of a G-study is to compute for variance components' estimates associated with a universe of admissible observations (Brennan, 2001a).

D-study

Generalizability theory (GT) separates decision study (D-study) from generalizability study (G-study). D-study focuses on finding strategies in reducing measurement errors investigate made for a particular purpose by the use of the information gotten from G-study (Brennan, 2001a; Crocker & Algina, 1986; Shavelson & Webb, 1991). D-study focuses on the computation, utilisation, and interpretation of variance components for making a decision with well-designed measurement approaches. Conceivably, the most significant D-study consideration is the description of a universe of generalization, which is described as a universe to which the investigator wishes to generalise depending on the outcome of a specific measurement process.

Types of GT Designs

One-Facet Design

From the perspective of GT, measurement is a sample from a universe of admissible observations, observations that an investigator is ready to use as a substitute for decision-making purposes. A one-facet design can be described

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

as a design with one source of error of measurement, that is, by a single facet. If the investigator (who is the decision-maker) intends to generalise from a single set of measurement items to a much bigger set of measurement items, items become a facet; the item universe would be defined by all admissible items. This suggests that if the main sources of measurement error in a particular study are two, then the design is called a two-facet design. Also, the design is called a three-facet design if the main potential sources of error of measurement are three.

Taking, for example, a students' appraisal of teaching exercise, where the only potential source of error is rater, there will be four sources of variability. One source of variability arises from systematic differences in instructors' performance. This facet is called the object of measurement. In practice, the object of measurement is not considered a source of measurement error. The second source, theoretically, is the rater facet. The third is the interaction between raters and the object of measurement. The last source of variation may rise out of chance, other systematic but unknown sources or unidentified sources of variation. In practice, the last two facets are merged into a single facet.

Designs with crossed and nested facets

In measurements in which all conditions of one facet (e.g. items) are observed with all conditions of another source of variation (e.g., raters) is known as a crossed measurement. One facet is said to be nested within another facet when more than one conditions of the nesting facets (raters) appear with one and only one condition of another facet (occasion). Two or more raters are present; on one occasion a different set of raters are present on another facet.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Random and Fixed Facets

Samples are deemed "random" when the sample size is much lesser than the universe size, and the universe size and the sample are either drawn randomly or considered to be interchangeable with any other sample of the same size taken from the universe. In other cases, however, the conditions of the facet exhaust the conditions in the universe to which the decision-maker wishes to generalise. In other cases, however, the conditions of the facet exhaust the conditions in the universe to which the decision-maker wishes to generalise. Consequently, such a facet is treated as a fixed facet.

Many-facet Rasch modelling

The MFRM originated from the traditional Rasch models (e.g., partial credit model, polytomous model, dichotomous model) that integrates several facets in the measurement processes (Bond & Fox, 2015). From the basic Rasch model, persons and tasks/items are assessed and positioned on an equal-interval scale with regards to their different item difficulty or the ability of the persons. The outcome of such estimations is sample independent. This is to say that, estimates of person abilities are not reliant on the particular sample of tasks/items used, and the item difficulties are equally not dependent on a particular group of individuals to whom the test was administered. The MFRM is an extension of the traditional Rasch model but comprises calculations on the influence of other sources of systematic error in the scoring which include rater discrepancies, differences in ratings due to occasions, and differences in the relative difficulty of the assessment tasks or items (Linacre & Wright, 2002).

According to Linacre (2013), a facet of measurement is a characteristic of the measurement process which the investigator believes can influence affect

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

test scores, and hereafter must be explored (Linacre, 2013). For instance, the rater severity, item or task difficulty, and rating scale category options can all affect the judgement of the rater. All estimations of the measurement facets are standardised on one equal-interval scale (usually called the logit scale), in so doing designing a single frame of reference for the interpretation of the data analysed.

Facets are concurrently estimated so they may be separately inspected. Most prominently, the MFRM offers information about how well the ability of each person/examinee, task, or rater is consistent with the expected estimates predicted by the strict arithmetic model produced during the data analyses. Consequently, MFRM can assist investigators to identify specific patterns within any facet that are "misfit" (i.e., the discrepancy between the observed and expected model). When there is a misfit, it denotes that the rater is unsystematically unpredictable in the application of the ratings, an item/task that is unpredictably difficult, or an individual whose responses are unreliable (Lynch & McNamara, 1998). In the framework of the MFRM, the fit indices are computed from the person/item residuals and are replicated in the outfit and infit mean square values with their expected estimate of 1.0 (Bond & Fox, 2015).

Due to the unique merits associated with the use of MFRM, it has been broadly applied in the area of health sciences, psychological and educational measurement, and language assessment (Bond & Fox, 2015; McNamara, 1996; McNamara & Knoch, 2012). In the area of language testing, MFRM is classically applied to rater-mediated performance assessments which include writing or speaking assessments where a score is the outcome of the interplay

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

between the items/task, the rater, the examinee, and the criteria (Batty, 2015). Particularly, this analytic procedure has shaped the foundation of the descriptor scales projected by the Common European Framework of Reference (CEFR) (North & Jones, 2009; North, 2000).

In addition, the MFRM model allows for bias analyses or interactions among the facets. For instance, it can specify whether a rater is inclined to rate a particular group of examinees differently than all others, or whether a particular group of examinees performed systematically differently on a task than they did on the other. Fundamentally, this is analyses of the difference in facet functioning, which is called differential item functioning (DIF) when the facet is item. The MFRM analysis of bias can also examine differential rater functioning (DRF), differential person functioning (DPF), and others based on the facet in question. Each interaction between the facets stated in the analysis is offered a biased score on the equal-interval scale, and its essence demonstrated by a standard z-score. These values with an absolute estimate which is equal to or larger than 2.0 designate significant interactions among the facets.

Differentiating between GT and MFRM

The application of GT thrives on the analysis of variance approach and this suggests that the data points were measured on an interval scale, meanwhile, these rating (or data) have a higher probability of being measured on an ordinal scale of measurement. Furthermore, G-theory emphasizes the use of the total rating or score of an examinee as the unit of analysis and presents this rating in the ordinal system of measurement of the original scores. In the MFRM analysis, on the other hand, the measures (rater severity, item difficulty,

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

examinee proficiency, among others) have the characteristics of a linear, equal-interval measurement scale if there is a data model fit. Consequently, these measures are appropriate for the computation of variances and means (Eckes, 2015).

Moreover, there is no adjustment in the observed scores in GT analysis for rater severity differences, discrepancies in criterion difficulty, or variances in measures linked to another facet. The purpose of GT is to calculate the error variance related to the raw scores of examinees with the goal of offering information on improving the dependability of measurements. The adjustment of score component in MFRM analysis offers linear measures regarding the examinee ability adjusted for the rater severity for specific examinee during the assessment/evaluation process (Eckes, 2015).

Moreover, the framework of GT focuses on the homogeneity of raters with the aim of allowing raters to operate such that they become interchangeable. In the case of MFRM, rater self-consistency is encouraged and, to some degree, expects raters to disagree with each other. These discrepancies in focus suggest a diverse conceptualisation of who an "ideal rater" is. From the viewpoint of GT, an "ideal rater" offers ratings devoid of any biases and errors, utilising the rating scale in an undistinguishable way under identical situations. Unlike the GT, the MFRM proposes that an ideal rater serves as an autonomous professional, demonstrating a particular degree of severity or leniency, and utilising the rating scale in a dependable way across criteria, items, examinees, among others. In the GT framework, between-rater severity variances are deemed as a cause of measurement error that needs to be minimised; advocates

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

of the MFRM method believe this to be a critical element of indelible variability among raters which only needs psychometric adjustment.

Although it is commonly approved that GT and MFRM modelling have a noticeably different methodology to address the problem of measurement error, scholars have also highlighted that these two modelling approaches have complementary usefulness (e.g., Iramaneerat, Smith, & Smith, 2008; Kim & Wilson, 2009; Lynch & McNamara, 1998; Sudweeks et al., 2005). In precise terms, this denotes the level of analysis describing each methodology: while GT sees the data principally from a group-level viewpoint, separating the sources of measurement error and calculating their extent, the MFRM analysis largely concentrates on individual-level information and, therefore, encourages functional examination into the functioning, or behaviour, of every individual component of the facets being considered (Linacre, 2001).

Therefore, there are some advantages in merging both GT and MFRM methodology dependent on the purpose of the evaluation or assessment. Kim and Wilson (2009) postulated that: If it is imperative to calculate the equivalence between the observed scores of the group of persons (examinee) and the raw scores that identical groups of examinees may receive under similar conditions, GT is relevant. If it is imperative to compute for each examinee a measure as free as imaginable of the particularities of the facets that produced the raw score, then MFRM is extremely recommended.

Identical Concepts Used in GT and MFRM

The comparison of GT and MFRM brings to bear three key concepts which are employed in the analytical framework of both approaches but have slight differences in their meanings for each of the methods. These concepts are

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

facet, interaction and reliability. To start with, the term facet in GT is described as a source of systematic error variations. Because variances produced by the object of measurement is deemed appropriate, facets in the context of GT comprise those elements of the model (with the exception of the object of measurement) which contributes unwanted variations to the general variability in scores. For instance, if a model is proposed and it consists of persons, raters, and tasks, GT sees persons as the object of measurement, and raters and tasks will be deemed as facets. MFRM, nevertheless, would describe all three components as facets. (Sudweeks et al., 2005).

The concept called "interaction" is also employed in GT and MFRM but used in somewhat diverse ways. Interaction in MFRM also known as bias analysis is explained as the interaction among facets and this demonstrates differential item functioning, differential rater functioning, differential person functioning, etc. These interactions are presented independently in a case-by-case manner. For GT, interaction is described as a form of factorial analysis in ANOVA. The analytical framework presents a variance contribution for every interaction of two or more major facets. The relative magnitude of these interaction variance elements shows the degree to which the comparative standing of the components in a single facet (or object of measurement) varies based on the components in a second facet. For instance, the variance component depicting the main effect for raters (σ_r^2) explains the degree to which raters in the data differ in terms of severity. Thus, the variance component for the interaction of persons and raters (σ_{pr}^2) explains the degree to which persons were differently rated by different raters (Shavelson & Webb, 1991).

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Lastly, the concept of reliability is utilised slightly different in GT and MFRM. In GT, the phi-and g-coefficients are presented as reliability or dependability coefficients and they are described as the reliability coefficient in the domain sampling model milieu. These reliability coefficients in GT explains how well the mean rating for a specific object of measurement predicts her universe score – the mean rating show could have received if rated on all probable prompts by all probable raters on any condition, where the domain criteria are fulfilled in each occasion. The g-index for a one-facet model is the same as Cronbach's alpha and may be understood in the same sense. This g-index may have a low value for G-studies that have more than one facet. The phi-index is naturally lower than the g-index because by definition it involves more errors or source of variance (Brennan, 2001b).

In contrast, the analysis from MFRM has two separate reliability indices for each facet utilised in research: (a) reliability of separation index, and (b) separation ratio. Even though these two reliability indices are presented in a different metrical scale, they are calculated from similar information and results to analogous inferences for a specific facet. Both indices define the degree of spread (or variability) in the measures computed by the MFRM for the several elements in the stated facet comparative to the precision by which those measures are computed. The reliability of the separation index for each facet is an estimate which varies between zero and one while the separation ratio varies from one to infinitude (Eckes, 2015). Both indices are reported for each major facet utilised in the MFRM analysis. Nevertheless, the functional meaning of these indices differs from facet to facet.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Higher estimates of the reliability of separation ratio and the separation index are appropriate for the person facet, but lower estimates are typically wanted for the other facets since variation among the components of these other facets results in 'construct irrelevant variance' in assessments. For the person facet, the reliability of the separation index is equivalent to Cronbach's alpha coefficient (Myford & Wolfe, 2003). The reliability of the separation index signifies the fraction of true score variance to observed score variance in the examinee ability estimates. In this research, the reliability of the separation index represents how well the raters and rating items differentiate among the quality of teaching course contents of the various lecturers (Bond & Fox, 2001). Estimates near 1.0 are appropriate for this index and demonstrate high reliability levels.

In cases where the person separation ratio equates to 1.0, the variability among the lecturers' quality of teaching and courses measures is no larger than the imprecision in the estimations of these measures. Therefore, there are no reliable means to differentiate the proficiency of one lecturer from one to the other. Nevertheless, if the separation equates to 2.0, then, the spread of lecturers' proficiency measures are two times the root mean square error, and if the separation ratio equates to 4.0 then the dispersions in the lecturers' measure are four times larger than the inaccuracy in their estimations (Fisher, 1992; Smith, 2001). Low estimates in either of these two values for the lecturer facet may signify central tendency error in the scoring, suggesting that the raters (i.e., students) were inept to differentiate between the performance of their respective lecturers on the activity performed (Myford & Wolfe, 2003, 2004). Lower estimates for both reliability estimates for anyone of the facets (except the object

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

of measurement) shows a high extent of reliability (i.e., presence of consistency) in the measures for the several elements of that specific facet.

Theoretical Review

The theoretical review, unlike the theoretical framework, stresses on the various theories that helps to explain the issue of investigation. Whereas the theoretical framework is the engine of the work (i.e., defines the objectives, the research design and analysis), the theoretical review only discusses why the need for the study (i.e., explain the issue of investigation). Three theories were reviewed: validity theory, Kirkpatrick evaluation model, and selection bias model.

Validity theory

Validity is a hypothetical term that has considerably changed some years now (American Psychological Association, 1954, 1966; American Psychological Association (APA), American Educational Research Association (AERA), and National Council on Measurement in Education (NCME), 1974, 1985, 1999). So many years back, validity was described as when a test measures what it purports to measure (Shepard, 1993) or a test was considered valid for anything with which it is related (Guilford, 1946). With time, these descriptions of validity were considered straightforward and non-problematic. Also, the term validity was seen as a fixed feature of a test. Sufficient relationships established test scores and some external measures were considered as a sign of validity.

Later, the straightforward description of validity changed and became more multifaceted, originally by adding different forms of validity evidence, namely, content validity, criterion validity, and construct validity. These diverse

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

validity forms were found to be associated with the test purpose. Content validity was utilised for tests that describe an examinees performance on a well-defined subject. Criterion-related validity was utilised for tests that predict examinees' future performance. Construct validity was applied in a situation where inferences needed to be made about psychological characters such as personality or intelligence. Cronbach and Meehl (1955) introduced construct validity as an indirect approach to validation which is used when no content domain or criterion variable can specify the extent to which a test measures what it is purported to measure.

A momentous revolution in validity theory was engineered by a statement by Cronbach (1971): 'one does not validate a test, but an interpretation of data arising from a specific procedure' (p.447). This implied that the focus of the validity examination drifted from the specific measurement tool to the meanings that can be made out of the measurement results. This viewpoint has persisted, and most scholars approve that it is not about the test itself but then how its results are interpreted and utilised in the validation procedure.

It must be mentioned that the diverse forms of validity evidence discussed in earlier paragraphs have been still relevant over the years but not the clear differences between them. The current validity theory has been known as the unitary validity framework, where construct validity has an overarching and central place by integrating virtually all types of validity evidence. This unitary concept of validity was progressively designed and eventually defined by Messick (1989), who recognised the agreement that constructs validity is the unifying concept of validity.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

According to Messick (1989), validity is described as a unified evaluative judgement of the extent to which practical evidence and theoretical justifications corroborate the appropriateness and adequacy of actions and inferences based on test scores. A substantial component in the unitary notion of validity theory was the primer of social consequences of measurement results. Later, Messick presented the concept of construct under-representation and construct-irrelevant variance as significant in the explanation of this unitary idea.

It was not only Messick who presented a broader framework (Crooks, Kane, & Cohen, 1996; Cronbach, 1988; Kane, 1992), but his model of validity became the well-known and established among those who supported a wide-ranging validity standpoint, and this was later added in the chapter on validity in the *Standards for Educational and Psychological Testing* (APA, AERA, & NCME,1999), demonstrating a universal recognition of Messick's viewpoint.

Introducing an expanded validity concept, where consequences are involved, resulted in numerous challenging concerns for the practitioner. A strong debate evolved concerning the viability of such a method and also about whether social consequences should at all be included in the validity theory. For those scholars who advocated the inclusion of consequences in the validity theory (Messick, 1989; Moss, 1998; Shepard, 1997; Stobart, 2001), they believed that examining the social consequences of measurement is necessary and logical since consequences are directly connected to the aim of a measurement process.

Scholars from the other side of the argument (Popham, 1997), for example, contended that adding social consequences would make the concept

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

of validity clutter and henceforward result in lack of clarity and confusion. Mehrens (1997) stated a similar view and indicated that the psychometric environment should narrow the usage of the concept validity instead of extending it. These scholars (i.e., Popham, 1997; Mehrens, 1997) requested that their aim was not to underestimate the essence of consequences associated with measurements, but instead to detached two vital, but, from their view, dissimilar concepts. These viewpoints are maintained in literature. Borsboom, Mellenberg, and van Herden (2004) argued that validity is now multifaceted that the theoretically-focused scholars have the probability of getting lost in the complex restraints of validity framework, whereas the practically-focused scholars are not likely to derive a practical conceptual structure with applied implications from it. Nonetheless, the extended concept of validity (and the application of social consequences as well) has also been applauded for improving the quality of validity surveys. Messick's viewpoint has been contended to be valuable for deriving and organising questions connected to validity (Gersten & Baker, 2002; Wikström, 2006; Wolming, 1999).

The 4th edition of Educational Measurement which was published in 2006, also had a new validity chapter authored by Michael Kane. Kane's (2006) view was not a contradiction of Messick's chapter but highlights practical viability via an 'argumentative approach' to validity. In this view of validity, two forms of arguments are emphasised: (1) the interpretive argument and (2) the validity argument. The interpretive argument should be a description of the planned interpretations and utilisation of test outcomes, from hypothesis to conclusions and decisions. Kane stresses that the level of validity of the measurement instrument has to do with the precision of the interpretive

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

argument of the measurement procedure and whether the validation argument, which is the evaluation of this interpretive argument, validates that the interpretive argument is reasonable and coherent and that the assumptions are credible.

The argumentative approach is consistent with the current viewpoint on validity and stresses that validity has to do with how a test is (interpreted and) utilised, although it is reported from a slightly different standpoint. The key difference with regards to the content is that it puts much less premium on consequences and particularly, social consequences likened to Messick's validity descriptions. The core merit of the argumentative approach is that it offers real-world supervision on how to apportion research efforts and in evaluating the procedure in the validation effort (Kane, 2006). Nevertheless, in line with Messick's 'unified evaluative judgement', Kane also explains the complication of the validation procedure and proposes a unified concept.

The validity theory helps explain the dynamics in the dependability of responses from students regarding how they appraise courses and teaching. The quality of teaching and course contents are "constructs" and as such pieces of evidence should be obtained to understand the interpretation and use of the data obtained from the students. Thus, data from students' appraisal of teaching quality is highly valid when several pieces of evidence can be provided with regards to how the data were taken, the meaningfulness of the data and how the data were used (Nitko & Brookhart, 2014). Hence, the variability surrounding the items, raters, instructors, among others should be looked at. In this reflection, some possibilities should be highlighted for purposes of gathering validity evidence: (a) raters may not be accurate in their ratings such that the

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

scores given may not reflects the skills and abilities of the instructors or overall teaching quality; (b) the evaluation items may not be clear enough for raters to understand, and (c) students may rate other characteristics of the course and instructor other than actual psychological construct which should be rated. These issues, among others, are likely to influence the fairness (and for that matter the validity) of the evaluation exercise and the results thereof may not be a true reflection of the construct being measured.

Kirkpatrick evaluation model

The Kirkpatrick evaluation model was propounded by Kirkpatrick in the year 1959. The model was broadly studied as an aspect of its semi-centennial merriments (Kirkpatrick & Kayser-Kirkpatrick, 2014). The model comprises four stages of evaluation developed to assess the efficacy of workplace training. These four stages are reaction, learning, behaviour, and results. The model is a typical framework for evaluating the effectiveness of training in a business setting (Figure 2). Though new frameworks have been projected, the four-stage model of training evaluation remains the most common and frequently quoted (Arthur, Bennett, Edens, & Bell, 2003a; Salas & Cannon-Bowers, 2001).

Reaction dimension

The reaction dimension focuses on the perception of the individuals who are being trained which in this case were the students who were receiving training from the school (Kirkpatrick, 1959, 1976, 1996). Alliger, Tannenbaum, Bennett, Traver, and Shotland (1997) discussed the difference between learner's reports concerning how much they have appreciated the training (affective reactions) and how well they think they have accumulated knowledge (utility judgments) within the reaction dimension. This criterion in the context of higher

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

education is characterised by students' evaluations of teaching and by self-report concerning supposed educational benefit.

Several scientists have highlighted the absence of association between the reaction dimension and the other three dimensions (learning, behaviour and results) (Alliger et al., 1997; Dysvik & Martinsen, 2008). Meta-analytic research revealed no association between affective reactions and other dimensions, and the presence of a weak association between utility judgments and the other dimensions were found (Alliger et al., 1997). Yet, notwithstanding the fact that several scholars have cautioned against the utilisation of only reactions for learning assessment, the reaction level dimension continues to be the most frequently evaluated (Arthur et al., 2003a; Van Buren & Erskine, 2002). From the viewpoint of Arthur et al. (2003a), a significant and probable reason for the extensive utilisation of the reaction level criteria is the affluence of collection. Nevertheless, many scholars also approve that learning, behaviour and results dimensions should be measured to precisely assess training outcomes.

Learning dimension

The learning level measure is the criteria of the learning outcomes, characteristically evaluated through the use of numerous systems of knowledge tests, but also by abrupt after-training assessment of ability and skill exhibition in the context of the training (Alliger et al., 1997). In the school setting, pre and post-tests offer the greatest direct evaluation of learning, and such measures are classically adopted in institutions of higher education (Arthur, Tubre, Paul, & Edens, 2003b). Nevertheless, it is also probable to utilise samples of writing, speeches, performances, and other assessments appropriate for a specific class.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

education is characterised by students' evaluations of teaching and by self-report concerning supposed educational benefit.

Several scientists have highlighted the absence of association between the reaction dimension and the other three dimensions (learning, behaviour and results) (Alliger et al., 1997; Dysvik & Martinsen, 2008). Meta-analytic research revealed no association between affective reactions and other dimensions, and the presence of a weak association between utility judgments and the other dimensions were found (Alliger et al., 1997). Yet, notwithstanding the fact that several scholars have cautioned against the utilisation of only reactions for learning assessment, the reaction level dimension continues to be the most frequently evaluated (Arthur et al., 2003a; Van Buren & Erskine, 2002). From the viewpoint of Arthur et al. (2003a), a significant and probable reason for the extensive utilisation of the reaction level criteria is the affluence of collection. Nevertheless, many scholars also approve that learning, behaviour and results dimensions should be measured to precisely assess training outcomes.

Learning dimension

The learning level measure is the criteria of the learning outcomes, characteristically evaluated through the use of numerous systems of knowledge tests, but also by abrupt after-training assessment of ability and skill exhibition in the context of the training (Alliger et al., 1997). In the school setting, pre and post-tests offer the greatest direct evaluation of learning, and such measures are classically adopted in institutions of higher education (Arthur, Tubre, Paul, & Edens, 2003b). Nevertheless, it is also probable to utilise samples of writing, speeches, performances, and other assessments appropriate for a specific class.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Alliger et al. (1997) indicated specifying knowledge retention, immediate knowledge, and skill/behaviour exhibition assessed within training as sub-types of learning level measures. This notion, however, received relatively little attention and acceptance.

Behavioural dimension

Behavioural dimension is also known as transfer level criteria, a jargon proposed by Alliger et al. (1997). The behavioural criteria comprise assessing actual on-the-job competency and can be adopted to recognise the influence of training on the performance of work. In organisations, this dimension is archetypally conceptualised as supervisor ratings or objective performance indicators such as job outcome (Arthur et al., 2003a; Landy & Conte, 2007). In the context of higher education, this dimension may consist of the evidence of learner's utilisation of skills and knowledge learned in previously engaged classes, including creative productions or research projects, learning application throughout a practicum, and in other behaviours outside the settings where the original learning took place.

After training behaviour can also be conceptualised as civic or workplace behaviour. Halpern and Hakel (2003) in their discourse on university learning emphasised the essence of transference in education and why students need to be taught in a manner that prepares them for unpredictable life tests, beyond the boundaries of the classroom. Though the learning and behavioural level criteria are theoretically anticipated to be associated, studies have revealed a comparatively uncertain association between the two (Alliger et al., 1997; Arthur et al., 2003a). This is attributed archetypally to the notion that after

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

training setting may or may not offer chances for the learned skills or material to be exhibited (Arthur et al., 2003a). This possible limitation needs to be well-thought-out in the assessment instruments design, and the collection and interpretation of behavioural data.

Results dimension

The results level dimension is both extremely required and more difficult to assess. In the business context, the results level dimension is conceptualised by efficiency gains, improved consumer satisfaction, improved employee motivation after training, or growth in organisational productivity (Arthur et al., 2003a; Landy & Conte, 2007). The result level dimension is the most difficult to judge. The outcome of this criteria is used significantly less often than evaluation of any other dimension. Alliger et al. (1997) called for attention to the fact that organisational restraints considerably limit chances for the collection of data of the results level criteria and repeat that those who fund training programmes may have impracticable expectations. Structural, economic and social limitations momentously affect not only the collection of data but the actual products on the results (and to a certain degree of behavioural dimension).

The Kirkpatrick evaluation model applies to this research which focuses on examining the dynamics in students' appraisal of courses and teaching (see Figure 2). Consequently, it can be stipulated that students evaluate courses and teaching based on the four indicators highlighted by Kirkpatrick (1959). Thus, if students enjoy the course the instructor is handling, and as well believe that they have learned much, then they will give a good rating (reaction dimension). Also, students will rate instructors excellently if they can demonstrate what they

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

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. Note that this level is beyond just writing and passing a test. Lastly, the result dimension comes in at the point where the students benefit from the course in the long term. The last dimension, which is the result, do not directly feature in this study since the ratings are done as at the time they are still in the classroom.

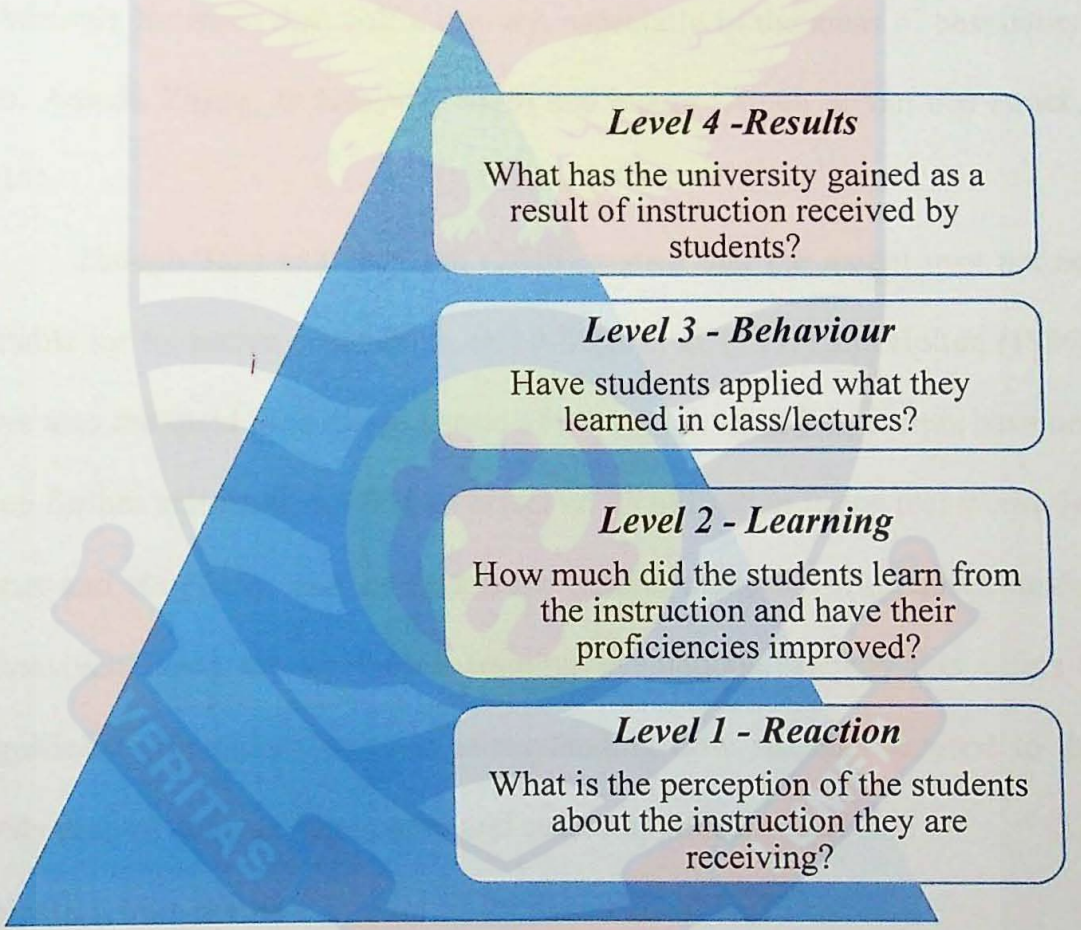


Figure 2- *Kirkpatrick Evaluation Model by Kirkpatrick (1959)*

The Kirkpatrick Model has been widely applied to the context of higher education with diverse viewpoints about its effectiveness (Abdulghani et al., 2014; Arthur et al., 2003b; Chang & Chen, 2014; Collins, Smith, & Hannon, 2006; Praslova, 2010; Paull, Whisted, & Girardi, 2016; Yardley & Dornan, 2012). There are pieces of evidence available to buttress the point that there is

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

a tendency for an evaluation exercise to be limited to the lower levels of this model (Steele, Mulhearn, Medeiros, Watts, Connelly, & Mumford, 2016).

The Kirkpatrick Model is a well-known and established method of evaluation that offers a robust edifice and does not need excessive time to be applied. Even though the model has its opponents and is not the only technique to assess interventions, the influence of the Kirkpatrick model in institutions cannot be underrated (Saks & Haccoun, 2010), given its extensive application in business for more than half a century, especially in the areas of hospitality (Ho, Arendt, Zheng, & Hanisch, 2016) and health (Ameh & van den Broek, 2015).

Though Saks and Haccoun (2010) argued that the model may not be suitable for formative evaluations, and Alliger et al. (1997) and Holton (1996) have also critiqued the ordered nature of the model, these judgements have not been further validated, nor had an effect on its utilisation in the real world. Its focus and simplicity, and its systematic nature, make the Kirkpatrick model extensively used for workplace training evaluations. Hence, this offers a significant preparatory point to understanding how students respond to the evaluation of the teaching quality and course appraisal.

Selection bias model

The selection bias model/framework was developed by Heckman (1974) and Gronau (1974). Whereas Heckman (1974) propounded the model and applied it to the supply of labour, market wages, and shadow prices in econometrics, Gronau (1974) applied this model to comparisons of wages in the labour market. The ideas of these two authors were all converged to the

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

selection bias model. Since then, the applicability of the model has expanded across several statistics related disciplines.

According to Heckman (1974) and Gronau (1974), selection bias occurs when observations about the outcome of interest are tied to some subpopulation that is non-randomly selected. In most cases, the characteristics of the subpopulation may go unnoticed or may perhaps be noticed after the outcome of interest has been observed. This translates into the problem of missing-not-at-random (MNAR) (Allison, 2001; Berg, 2005; Rubin & Little, 1987). This model underscores that non-response, particularly missing-not-at-random, results in selection bias because sub-groups may go unnoticed. In the case of students' evaluation of courses and instructors, selection bias is observed when evaluation results are taken from student evaluators who were not selected randomly on observable factors but rather selection was done based on some evaluators' latent traits such as evaluators' views of anonymity of data or motivation to respond to the instrument.

The decision of student evaluators to whether participate or not to evaluation survey is based on their net utility derived from the response. That is, the satisfaction students derive from responding to course and instructor evaluation survey reinforces their decision to respond. This satisfaction is a function of covariates (with their coefficients) plus an added error term. Student evaluators who do not derive satisfaction from responding to the evaluation are unlikely to respond to the evaluation surveys. The opposite is right for students who derive satisfaction in responding to the evaluation. The difficulty in ascertaining an estimate of this net utility impedes knowing which category of students are unwilling to participate in the evaluation of courses and teaching.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

However, there are some indicators to predict the likelihood that student evaluators will or will not respond to an evaluation instrument (Heckman, 1974).

The observable indicators which manifest as covariates to drive students' responses to evaluation have been extensively discussed in the literature. The prevailing literature on students evaluation of the course and teaching have found that students who are high-achievers (using cumulative grade point average or grade) and, in utmost settings, female students have a greater probability of responding to the evaluation of courses and instructors (Adams & Umbach, 2012; Avery, Bryant, Mathios, Kang, & Bell, 2006; Kherfi, 2011; Marsh, 2007; Spooren & Van Loon, 2012; Porter & Umbach, 2006; Porter & Whitcomb, 2005). Adams and Umbach (2012), in their observations, reiterated that if students do not desire to remember a course in which they performed poorly, there is a low probability of them responding to the evaluation. The majority of evaluations are done in the latter part of the semester and thus, students might have received their scores from the quizzes or mid-term examinations already taken. I agree with the other scholars that this might influence students' decisions to or not to respond to the evaluation.

Additionally, survey fatigue has been found to reduce responses to the evaluation of courses and instructors, where the cause of fatigue is as a result of responding to many surveys surrounding similar issue simultaneously and/or lengthy evaluation items (Adams & Umbach, 2012; Groves, Fowler, Couper, Lepkowski, Singer, & Tourangeau, 2004; Spooren & Van Loon, 2012). In reality, this might influence the evaluation data due to the way evaluation is done. Speaking from my observations from universities in Ghana, evaluation of

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

courses and instructors in most universities are done around the same time (i.e. getting to the end of the semester) whether it's online or paper-based. By implication, students taking six courses, for instance, are required to respond to the evaluation instrument six times at approximately the same time. Student evaluators might end up not responding at all to the evaluation or skip some of the items or in the worst case provide inaccurate ratings.

While some of the indicators influencing students' response to evaluation are observed, not all are observed in practice. Students' standards of learning behaviours (i.e. socialization and their intrinsic motivation for education), for instance, are likely to play a significant role in students responding to evaluation (Kherfi, 2011; Wolbring & Treischl, 2015). Precisely, students with great intrinsic motivation might feel more participated in the process of educating, thereby, increasing their rate of response. Lastly, student evaluators may also unlikely to respond if they are apprehensive about evaluation anonymity (Dommeyer, Baum, Hanna, & Chapman, 2004): these are largely unobserved.

A typical conditional, or observed, instructor or course appraisal score, is prejudiced if it varies from its unconditional equivalent (i.e. the mean score of the evaluation data assuming every student participated). To create an accurate picture, if students involved in the evaluation of the course and instructor, does the evaluation more leniently for some known/unknown reasons, the mean observed evaluation score is likely to be biased leading to higher evaluation scores. It is clear from the literature that several indicators (whether observed or unobserved) influence the students' response to

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

evaluation (Adams & Umbach, 2012; Dommeyer et al., 2004; Groves et al., 2004; Kherfi, 2011; Spooren & Van Loon, 2012; Wolbring & Treischl, 2015).

Students' evaluation of courses and instructors mostly suffer from low response rates, especially when the evaluation is done for all courses at approximately the same time and/or done online as is seen in universities in Ghana and beyond. If this happens to be the situation, evaluation scores would be misrepresentative of all the students' opinion regarding the evaluation objects. This is because some subgroups in several population and sample with similar characteristics are found to have a higher probability of responding to evaluation.

Suppose that a large majority of high-achieving students respond to the evaluation of their instructor, there is a potential selection biased even though the investigator might not be aware of it. In a situation like this, the evaluation data will reflect the opinions of high-achieving students rather than the views of the entire class. This has implication for the validity of the data taken. In other words, the results may be distorted limiting the soundness of the interpretations and the inferences made from the data (a measure of teaching quality). In fact, a comparison of scores from evaluation of instructors with similar quality of teaching but different rates of response has shown a high evaluation score for instructors with a low rate of response (Goos & Salomons, 2017). Goos and Salomons further cautioned the use of evaluation data for decision making when there are low response rates across different courses but comparisons should be made based on the similarity of response rate.

In relation to this current study, it is obvious that the response rate is a quality indicator of evaluation data taken from students. Several pieces of

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

evidence have been provided by numerous scholars in explaining how the response rate can affect the validity of evaluation data. It is only imperative that in examining the validity of students' evaluation of courses and instructors, the response rate should be factored and used as a covariate. It must be emphasised that the prevalence of response rate does not only exist in theory but in practice. It is real in most cases, if not in all cases, that data taken through the online survey suffer non-response. This study is not an exception since the evaluation was done in the form of a survey by the university. This study highlighted whether this non-response can affect the validity of students' evaluation data as stipulated by literature.

Empirical Review

The empirical review comprised related studies which attempted to investigate the validity of students' appraisal of courses, teaching and learning. For clarity, the review was organised based on the approaches (CMT, GT and MFRM) adopted by the previous studies. This was because these methodological approaches were the centre of the investigation in this research and thus, this study's objectives and research questions were embedded in that methodology.

Studies that employed CMT as the methodology

To start with, Samian and Noor (2012), in their study, explored the reliability of students rating of lecturers' performance assessment at Universiti Teknologi Malaysia. A survey was conducted to gather data on the students rating of the lecturer's performance assessment (teaching). Since it has been a norm for that university to engage students in the evaluation of lecturers' teaching, a standardised instrument used by the university was adopted by

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

In a similar study, Sok-Foon et al. (2012), examined students' evaluation of instructors' performance in a private university in Malaysia. The authors were motivated by the fact that there were predictors or factors which determines the evaluations of students regarding their lecturers' teaching. To identify these factors, a survey was conducted among 223 student evaluators enrolled in undergraduate programmes in the university. An existing questionnaire that the university used was adopted and used for their data collection. A multi-stage sampling technique was employed in sampling the students. It was revealed that the sources of variances of student evaluation were students' characteristics, instructor characteristics, learning facilities and resources, and studentship (i.e. students' commitment to education). Together, these factors contributed about 61.9 percent of the variances in students' evaluation of instructors' teaching performance. The approach used by these authors makes it difficult to disentangle the sources of variances, thereby, obtain the exact amount of variance for each factor. Again, the authors were silent on the number of classes or lecturers evaluated. This makes the results unclear since every student rating is based on the extent to which the teacher performs. Aggregating the responses becomes problematic at this stage.

In another study that was similar to that of Samian and Noor (2012), Zhang et al. (2015) identified the qualities of teachers that university students sought to look out for. Their study, implicitly, examined four sources of variance in students' evaluation of teachers: gender of student evaluator, perception of students, characteristics of teachers, and university evaluators attend. Only economics students from the two universities were sampled for the study. A cross-sectional survey was conducted and the ratings of the students

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

(both undergraduate and postgraduate) were observed for four semesters in the first university (South-West United States), and a period of two semesters for the second university (South-East United States). An instrument developed, validated and used in one of the universities was used for the study. A total of 458 students were sampled from the two universities for the study. Pearson Product Moment and Spearman Rank correlation were used for the analysis. It was revealed that the students were consistent in their ratings regarding the qualities they seek in a teacher. Male and female students differed in their ratings indicating that the variance for gender was likely to be high. Students from the two universities rated differently showing a high proportion of variances in the universities.

From the findings of Zhang et al. (2015), the differences in the ratings of the students in the two universities, from my view (the researcher of this present study), could be attributed to the fact that the questionnaire used was adapted from one of the universities. It is possible that students' familiarity (i.e., students who have previously responded to the instrument) or unfamiliarity (i.e., students who have never previously responded to the instrument) factor can ignite variations in the responses. Notwithstanding, some characteristics of the specific schools such as policies, school climate, and teacher-student relationship, can also play a significant role. The approach employed by Zhang et al. (2015) had a lot of flaws and this cast doubts regarding the accuracy of their conclusions made.

Fah and Osman (2011) examined the factors influencing students' evaluation of university lecturers in a selected university in Malaysia. The authors conveniently sampled the course and their respective instructors. In all,

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

88 students enrolled on an undergraduate programme were sampled and questionnaires were administered to them through a survey. The sources of variances which their study focused on were instructor characteristics, tutorial ratings and course characteristics. The instrument for data collection was developed and validated by the authors of the study (Fah & Osman, 2011). The study found a high rate of consistency in the ratings of students regarding instructor characteristics, tutorial ratings and course characteristics. Most importantly, these indicators were found as the sources of variances to students' evaluation with the largest variance coming from the instructor characteristics, followed by the characteristics of course. Tutorial ratings did not contribute any variance to the ratings of students. Using the CMT approach has some implications for the study. This raises concern about the validity of the results. Again, the number of student evaluators used in Fah and Osman's study were relatively few to assure sufficient generalisation.

Üstünlüoğlu and Can (2012) investigated the area of student evaluation of their teachers. The study was conducted to examine if there is any relationship between students' evaluation of teachers and coordinators evaluation of teachers and explore the agreement in rating across two years. Conducting the study at English Medium University, teachers teaching at a preparatory English programme in Turkey were engaged to participate in the study. Data were taken in two academic years. In the first year 1,028 students, 99 instructors and 9 programme coordinators participated in the study. For the second year, 1,211 students, 99 instructors and 9 programmes coordinated were involved. Üstünlüoğlu and Can developed and validated a questionnaire for data collection. Results from the Pearson Product Moment Correlation revealed a

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

that even in a milieu where the curricula are fixed and all instructors in the same course taught with precisely identical materials, instructors still matter significantly indicating those instructor characteristics had a high variance proportion. The mean difference in successive achievement between students who were allocated to the worst and best instructor (on the effectiveness scale) is about 43 percent in the exam grades distribution, equivalent to approximately 5.6 percent of the mean grade. Furthermore, it was found that instructors with better successive achievement received poor students' ratings indicating large variances in student evaluators. It was again found that classes with a greater number of high achieving students produced stringent evaluations and this may affect the validity of the results.

In a study that was similar to that of Braga et al. (2011), Raza and Irfan (2018) assessed the students' ratings of their instructors' attributes. The study was purely quantitative. Through a cross-sectional survey, 960 students were sampled through stratification from the public (4) and private (6) universities in Pakistan. About 30 student evaluators were sampled from each department using a systematic sampling technique. A multi-dimensional standardised scale developed and validated by one of the authors was adapted for the study. One-way ANOVA, Pearson Product Moment correlation and independent samples t-test were used to analyse the data. The study found instructor characteristics as the major source of variations in the ratings of students. There was a high rate of consistency in the ratings of students regarding their instructors. The programme of study of the students also had a large variance proportion to the ratings of students. The variance was, however, low for the type of university. The CMT approach was used.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

The studies conducted discussed so far (Braga et al., 2011; Raza & Irfan, 2018; Fah & Osman, 2011) only explore the sources of variances in students' evaluation of instructors teaching or course evaluation. Many of these studies only estimated the composite number of variances of all the factors they explored. In a very few instances, that some studies examined the separate proportions of variances of each source of variability. These studies failed to delve so much into how accurate and reliable the data were. All the studies stayed glued to the use of CMT approaches which is group dependent and comes with their flaws.

Unlike the previously reviewed studies, other pieces of research examined the extent of non-responses in students' appraisal of teaching and how these non-responses affected the accuracy of responses provided by the students. Such studies investigated how non-response can serve as a source of variability in students' evaluation of teaching. Reisenwitz (2015), for example, examined the non-response bias in student online evaluation of instruction. The study assessed the differences in students' evaluation of instruction between students who completed the online evaluation form and those who failed to participate in the evaluation of their teachers. About 373 undergraduate students were sampled from a university in the South-Eastern United States. The university's existing system of evaluation of instruction- online students' evaluation of instruction- was adopted for the study. Besides, the traditional mode of students' evaluation of teaching (paper-and-pencil form) was also utilised. All students who participated in the study were sampled from multiple business classes.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Goos and Salomons (2017) also assessed selection unfairness in the evaluation of courses through the measurement of teaching quality in higher education. Their research aimed at exploring how survey non-response rate and non-random selection of student evaluators can influence the evaluation data. The authors initially attempted the trend and magnitude of selection bias in evaluation which results from both unobserved and observed characteristics of more than 3000 taught courses in a very populated university in Europe. The data were taken from the web. The data were obtained from programme areas including Bio-Medical Sciences, Technology and Science, Engineering, and Social Sciences and Humanities. An already existing instrument used by the university was adopted and used. It was found that students' evaluation of courses in the university was increasingly biased and that when this biased was corrected through an econometric model, a significant change in the average rating score and the ranking of evaluated courses was revealed. The findings from their study indicated that students' evaluation of instructors' quality of teaching somewhat cannot be relied upon.

Similarly, Reisenwitz (2015) found that non-response bias continues to prevail in student evaluation of teaching, especially with an online survey. Results from the t-test and logistic regression revealed a statistically significant difference between students who responded to the evaluation and those who failed to complete it. Further findings revealed that race, gender, and cumulative grade point average were the factors influencing non-response of students' evaluation of instruction. Although Reisenwitz's finding provides some clue about the effect of non-response, issues are surrounding how the data were obtained for those who responded to the online survey and those students who

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

failed to respond. The data for these two groups were obtained on different occasions and this might have contributed to the observed differences between the two groups.

Adams and Umbach (2012) conducted a study that was similar to that of Reisenwitz (2015). Adams and Umbach explored the rate of non-response in online students' evaluation of instruction. Their study further examined the factors influencing the involvement of students in evaluating their lecturers at North Carolina State University. Data were taken from 134,929 undergraduate over a 3-week duration. The survey was conducted online. The web-based evaluation system utilised by the university was adopted for their study. Similarly, the questionnaire developed and validated by the institution was also adopted by the researchers. The study revealed non-response rates between 51 and 39 percent. Their study suggested that there is a higher likelihood that the non-responses may influence the outcome of the evaluation. Factors such as fatigue and salience were found to influence the non-participation of evaluation by students.

Although Adams and Umbach (2012) concluded that non-response was critical to the quality of data provided by the students, it was unclear why such a conclusion was made. This was because there was no concrete approach used to establish this relationship between non-response and the accuracy of the data provided by the students. What Adams and Umbach just did was to establish whether non-response was present and what factors led to students not responding to the evaluation survey.

Sliusarenko (2013) investigated the validity and sources of variability to students' appraisal of teaching by emphasising two factors: occasion and non-

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

response. This notwithstanding, the overarching aim of Sliusarenko's research was to develop effective procedures for analysing data from student evaluation of courses and instruction at the Technical University of Denmark. The data were taken from two intervals: the middle of the semester and the end-of-semester. About 4, 211 students were sampled from various programmes at the Diploma, Bachelor and Masters level of education. The author employed classical and modern approaches to multivariate statistics to analyse the data. It was found that student evaluation of instructions conducted in the middle of the semester appeared more dependable than the one conducted at the end of the semester. The items utilised were found to have larger variances and such the need to reduce the number of items. There was evidence of non-response among the students.

The study by Sliusarenko (2013) has two major limitations. First, the reliability evidence found by Sliusarenko can only be attributed to the occasion within which the study was conducted. Other relevant proximal factors which include rater variability and item difficulty might have contributed to the nature of results found. Secondly, exploring whether non-response was present is not sufficient enough to understand how student non-response is linked to the accuracy of responses.

Studies that employed GT and MFRM as a methodology

Some studies were also conducted using the GT and MFRM approach. In a study conducted by VanLeeuwen, Dormody, and Seevers (1999), for example, GT was employed to evaluate the dependability of students' evaluation of lecturers teaching. Using a standardised and validated instrument on a 5-point scale, a survey was conducted among 531 students during the fall

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

season. Data were taken from the College of Agriculture and Home Economics at the New Mexico State University, New Mexico. Data were specifically taken from 30 classes through purposive sampling ensuring that all manner of courses were represented in the college. During the study, the authors ensured that lecturers or courses were not duplicated which prevented confounding results. The students sampled from each of the classes ranged from 7-24. Although not clearly stated in their study, a three-facet nested random design was used. These facets were items, classes, and students. The residual (interaction of all the facets) contributed the largest variances in the ratings of students. Their study revealed that students nested in classes (*s,sc*) contributed the second-largest variances to the responses of the students. The item facet contributed the smallest variances to the ratings of students.

VanLeeuwen et al. (1999) further modified the original five-point scale to a two-point scale by merging strongly agree and agree as a positive response, and merging neutral, disagree and strongly disagree as a negative response. Similar results were obtained when the alternative scoring approach was employed- 2-point response type (positive response and negative response). Results from the dependability and generalizability index revealed that a high level of reliability indicating that the data taken from students on the ratings of lecturers have a high degree of validity.

Feistauer and Richter (2016) asked a question on whether students' evaluation of teaching quality can be relied upon. This set the tone for their investigation. The authors used a variance approach to achieve the purpose of their study. The study used a three-year data set from 2011 to 2014 at the University of Kassel in Germany. Fifty-three (53) instructors were observed

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

during the three years while engaging in lectures and others offering seminars. The number of students who participated in the study were 480. With a unique identification, students were kept anonymous, while data from the same students were traced and matched. Courses included statistics, clinical psychology, social psychology, educational psychology and cognitive psychology. Both Bachelor of Science and Master of Science (Psychology) students participated in the study. A widely used questionnaire (5-point scale) for evaluation by German universities was used. The variance components were students, courses, teachers and students/teacher interaction.

Feistauer and Richter (2016) found that the considerable variance components of courses and teachers revealed some level of reliability. Nevertheless, a comparable percentage of variance was found to be contributed by students, and the interaction of teachers and students was the largest source of variance. That is, the raters' individual opinion of teaching and the alignment of these views with a specific teacher significantly influenced their ratings. The results suggested a flaw and cast some uncertainty on the validity of student rating of teaching quality and that caution should be taken when using the aggregated evaluation scores.

Quansah (2020) assessed the reliability of students' rating of lecturers teaching. The study explored: (a) sources of variances that contribute to the ratings of students, (b) the accuracy of students' rating of lecturers, (c) decisions to be taken to increase the dependability of students' evaluation. Coming from the positivists' paradigm and using a correlational research design, a three-facet partially nested (i.e., Rater, Occasion, and Evaluation Items) G study and Decision (D) study design were employed. Data were taken from two points in

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

time (middle of the semester and end-of-semester) using the students as raters, evaluating the lecturers on all the items. Twenty (20) students rated each of their lecturers. In all, 30 lecturers were evaluated. An already existing evaluation questionnaire used by the university was validated using Structural Equation Modelling and used thereafter for this study. Results from a univariate analysis revealed that the residual (interaction of all facets) contributed the largest variance to students' ratings. The results found that the overall dependability of students rating of lecturers was low. Further analysis revealed that evaluation in the middle of the semester produced a better outcome. The low dependability of students' rating is influenced by the evaluators (raters), evaluation items, and occasions of rating.

Li et al. (2018) explored the evaluations of teaching carried out by college students. By carrying out a survey, the authors sampled 566 college students from 19 classes within three universities in China. Data were obtained from the students using two data points: first data taken at the end of the fall (first) semester and the second was taken at the beginning of the spring (second) semester. The classes cut across several programmes of study including science and arts programmes. In all, 19 instructors were evaluated. The authors developed and validated their instrument for the study. Multivariate GT was used to analyse the data. The results showed that students' evaluation of instructors at the start of the second semester was more valid than evaluation from the first semester. Again, evaluation by college students reading science curriculum was found to be better than evaluation data from the liberal arts curriculum. Analysis from the decision study found that to ensure better dependability, 20 college students were sufficient to rate a single lecturer. Li et

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

al., however, failed to estimate the overall dependability and generalizability index for the data to tell whether the data taken from student evaluators could be trusted. Inferring from the data nevertheless, suggest that, the validity of the evaluation data should be questioned. Again, the authors neglected the significant role that rating items play in explaining variability in ratings.

Ogbonnaya (2019) assessed the reliability of students' evaluation of instruction at some selected secondary schools in South Africa. Using a quantitative approach, a survey was conducted among grade 11 students to evaluate the teaching of mathematics across 8 classes. The convenient sampling procedure was used to sample students in the selected classes to participate in the study. Mathematics teachers in the selected schools were the object of measurement. Data were taken from 194 students with the number of student evaluators from each class ranging from 17 to 34. A questionnaire (on 6 point scale) developed and validated by the researchers was used for the data collection. The data were analysed using an average deviation index and intraclass correlation coefficient. The study revealed a high point of reliability of the evaluation. Thus, there was high consistency and closely perfect alignment in the evaluations of the teachers over all of the evaluation items. The author concluded that secondary school students can offer a dependable evaluation of their teaching quality. Ogbonnaya's (2019) study used a CMT which is sample dependent. This approach is a weak way of estimating the reliability of psychological measures.

Unlike the previously reviewed studies, Börkan (2017) explored the sources of variability in students' evaluation of instructors teaching in a selected university in Turkey using MFRM. Börkan examined the extent to which

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

student, rating items, and instructor contributed to their ratings and as well explored students' judging behaviours. A survey was conducted using 1,235 students in 254 courses. A standardised evaluation questionnaire that was already used by the university was adopted. The least student who evaluated each course was 6. The Faculty of Science and Art was purposively sampled and students and courses within this faculty were used. The faculty was used because it offered courses to all students in the university. It was revealed that the students largely differed in the level of severity while rating instructors. Halo effect was prevalent at both group level and individual level. The variances in the rating item were found to be fairly low. The findings of the study questioned the validity of students' rating as students failed to evaluate teachers as was expected of them.

The studies which adopted GT had a number of flaws. These studies were sample dependent and the facets were inter-related. This implies that the results they had, especially with the optimisations, could change when a facet was either removed or another facet added. How then can the optimisation results be used to make decisions which should inform practice? Furthermore, these studies which adopted GT could not indicate the level of raters severity because the GT could not implement such analysis. Also, such studies failed to evaluate the functioning of the rating scales employed in their study. The weaknesses of the GT could have been supplemented with the use of MFRM. Unfortunately, no study was found to have combined but approach. The study which used the MFRM approach by Börkan (2017) did not add GT. There was, therefore, the need for a study that employed both approaches to understand the variability in students' evaluation of courses and teaching.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Table 1- *Summary of Empirical Studies*

Author(s)	Study Area	Approach	Facets	Instrument details	Findings
Samian & Noor (2012)	Universiti Teknologi Malaysia	CMT	Students	An existing instrument used by the institution; unidimensional	A high rate of correlation was found between the comments and the ratings. The study established that students rating of the lecturer's teaching quality was highly valid and reliable.
VanLeeuwen et al. (1999),	New Mexico State University, New Mexico.	GT (univariate)	Three facets (items, classes, and students)	Standardised questionnaire; unidimensional	The residual (interaction of all the facets) contributed the largest variances in the ratings of students. High level of validity of students rating
Feistauer & Richter (2016)	University of Kassel in Germany	Variance approach (GT)	Three facets (students, courses, and teachers)	An existing instrument used by the institution; multidimensional	The results suggested a flaw and cast some uncertainty on the validity of student rating of teaching quality and that caution should be taken when using the aggregated evaluation scores. The interaction of teachers and students was the largest source of variance.
Börkan (2017).	A selected university in Turkey	MFRM	Three facets (student, rating items, and instructor)	An existing instrument used by the institution; unidimensional	It was revealed that the students largely differed in the level of severity while rating instructors. The findings of the study questioned the validity of students' rating as students failed to evaluate teachers as was expected of them.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Table 1 (Continued)

Quansah (2020)	A selected university in Ghana	GT (univariate)	Three facets (student, rating items, and occasion)	An existing instrument used by the institution; unidimensional	Results from a univariate analysis revealed that the residual (interaction of all facets) contributed the largest variance to students' ratings. The results found that the overall dependability of students rating of lecturers was low. Further analysis revealed that evaluation in the middle of the semester produced a better outcome.
Li et al. (2018)	Three selected universities in China	GT (multivariate)	Four facets (Evaluation time, students, major type, and curriculum type)	Developed and validated instrument by the authors; unidimensional	The results showed that students' evaluation of instructors at the start of the second semester was more valid than evaluation from the first semester. Ratings of students appeared to be inconsistent across time and programme major
Goos & Salomons (2017)	A selected university from Europe	Econometric model	Response rate & selection bias	An already existing instrument used by the university: unidimensional	It was found that students' evaluation of courses in the university was increasingly biased. The findings from their study indicated that students' evaluation of instructors' quality of teaching somewhat cannot be relied upon.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Table 1 (Continued)

Ogbonnaya (2019)	Some selected secondary schools in South Africa	CMT (Average deviation index and intraclass correlation coefficient)	Students	Developed and validated instrument by the author; unidimensional	The study revealed a high point of reliability of the evaluation. Thus, there was high consistency and closely perfect alignment in the evaluations of the teachers over all of the evaluation items.
Üstünlüoğlu & Can (2012)	An English Medium University, Turkey	CMT (Pearson Product Moment Correlation)	Students	Developed and validated instrument by the author; unidimensional	Results revealed a statistically significant positive relationship between students' evaluation and coordinators ratings of the same instructor. A statistically significant positive association was found in the ratings of the students over the two-year period.
Sok-Foon et al. (2012)	A private university in Malaysia	CMT (regression)	Students	An already existing instrument used by the university; unidimensional	It was revealed that the sources of variances of student evaluation were students' characteristics, instructor characteristics, learning facilities and resources, and studentship (i.e. students' commitment to education). Together, these factors contributed about 61.9 percent of the variances in student evaluation of instructors' teaching performance.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Table 1 (Continued)

Zhang et al. (2015)	Two universities in US	CMT (Pearson Product Moment and Spearman Rank correlation)	Students	An instrument developed, validated and used in one of the universities in the study; unidimensional	It was revealed that the students were consistent in their ratings regarding the qualities they seek in an instructor. Male and female students differed in their ratings Students from the two universities rated differently
Fah & Osman (2011)	A selected university in Malaysia	CMT (regression)	Students	Developed and validated instrument by the authors; unidimensional	The study found a high rate of consistency in the ratings of students. The sources of variances to students' evaluation with the largest variance coming from the instructor characteristics, followed by the characteristics of course. Tutorial ratings did not contribute any variance to the ratings of students.
Braga et al. (2011)	Bocconi University in Italy	CMT (Finding differences in groups)	Students	An already existing instrument used by the university; Multi-dimensional	Furthermore, it was found that instructors who are related with better successive achievement received poor students' ratings indicating large variances in student evaluators. It was again found that classes with a greater number of high-achieving students, produced stringent evaluations and this may affect the validity of the results.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Table 1 (Continued)

Raza & Irfan (2018)	4 Public and 6 Private Universities in Pakistan	CMT (Findings differences in groups)	Students	A multi-dimensional standardised scale from developed and validated by one of the authors	The study found instructor characteristics as the major source of variations in the ratings of students. There was a high rate of consistency in the ratings of students regarding their instructors. The programme of study of the students also had a large variance proportion to the ratings of students.
Reisenwitz (2015)	A university in the South-eastern United States.	CMT	Students	An already existing instrument used by the university; unidimensional	Non-response bias continues to prevail in student evaluation of teaching, especially with an online survey. Results revealed a statistically significant difference between students who responded to the evaluation and those who failed to complete it. Race, gender, and cumulative grade point average were influenced by non-response of student evaluation of instruction.
Adams & Umbach (2012)	North Carolina State University, USA	CMT	Students	An already existing instrument used by the university; unidimensional	The study revealed a non-response rate between 51 and 39 percent. The study suggested that there is a higher likelihood that the non-responses may influence the outcome of the evaluation.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Table 1 (Continued)

Sliusarenko (2013)	Technical University of Denmark, Denmark	CMT	Students	An already existing instrument used by the university; unidimensional	It was found that student evaluation of instructions conducted in the middle of the semester appeared more dependable than the one conducted at the end of the semester. The items utilised were found to have larger variances and such the need to reduce the number of items. There was evidence of non-response among the students.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Conceptual Framework

The conceptual framework has four predictor variables (i.e., raters, lecturers, items and non-response) and a criterion variable which the validity of students' ratings of courses and teaching (Figure 3).

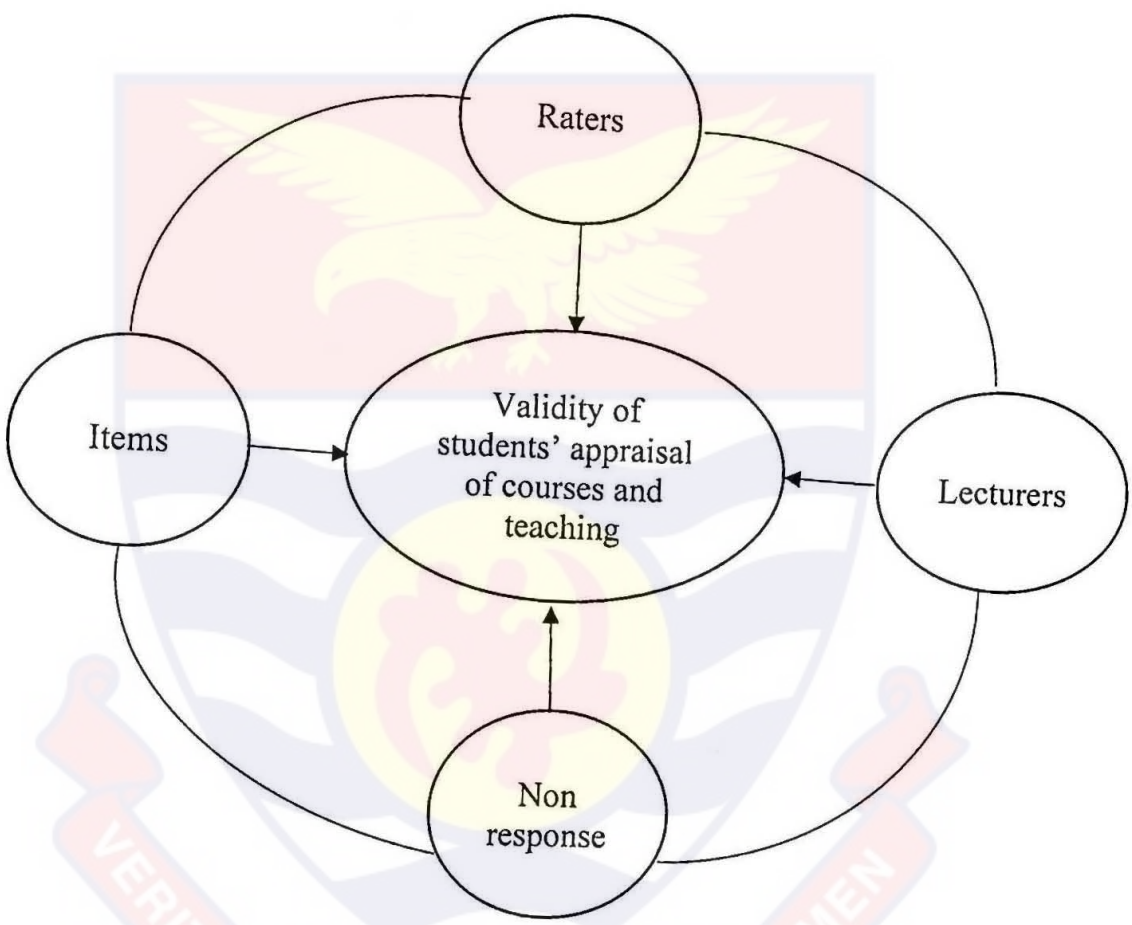


Figure 3- Conceptual Framework

As shown in Figure 3 (i.e., the conceptual framework), this study proposes that raters, lecturers, evaluating items, and non-response contributes to the variability in students' evaluation of courses and teaching. In practice, when students are requested to appraise courses and teaching quality, there is a chance of disagreement among and within themselves due to several factors (Eckes, 2015). Some of these factors are what this framework has outlined. It must be noted that giving a score to depict the degree to which a particular trait

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

is possessed by the object of measurement (i.e., instructor) is purely based on the subjective judgement of the students. Instead of obtaining a common score for instructors, student regularly seems to considerably differ in their ratings. This has been seen to be caused by the items used. Either the items are unclear, ambiguous, or the scale is unclear. In all the foregoing situation, the context (paper or online evaluation), as shown in the framework, also contribute to the errors in rating. This research sought to understand how these variables help explain the dependability of students' appraisal of courses and teaching.

Chapter Summary

This chapter extensively reviewed information that was relevant for the discussion of the issue under investigation. The literature review started with the explanation of key concepts on students' appraisal of teaching to provide grounds for the readers to understand the focal issues. This was discussed in connection with the key dimensions of the evaluation process: the teacher, teaching process and learning outcome. The conceptual review ended with the utilisation of students' appraisal data.

The review also highlighted the theoretical framework and theoretical review. The major two modelling theories (GT and MFRM) were used as the theoretical framework to provide a frame that guided the conduct of the study. The validity theory, Kirkpatrick evaluation model and selection bias model were used as the theoretical underpinning models for the study. How these frameworks and models were related to the research were outlined. Based on the framework the conceptual framework was designed.

A monumental component for the literature review which served as the engine of the work was the empirical review. The empirical review included

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

discussions on the literature on issues of validity and dependability of students' appraisal of courses and teaching around the globe. The empirical review made it clear that there are three predominant approaches to the investigation of the validity of evaluation data, which are CMT, GT and MFRM. From the review, the majority of these studies employed CMT which is the foundation of most, if not all, measurement theories (e.g., Adams & Umbach, 2012; Braga et al., 2011; Fah & Osman, 2011; Goos & Salomons, 2017; Ogbonnaya, 2019; Samian & Noor, 2012; Sliusarenko, 2013; Raza & Irfan, 2018; Reisenwitz, 2015; Sok-Foon et al., 2012; Üstünlüoğlu & Can, 2012; Zhang et al., 2015). Very few studies, on the other hand, used GT (e.g., Feistauer & Richter, 2016; Li et al., 2018; Quansah, 2020; VanLeeuwen et al., 1999). Only one study investigated the dependability of students' evaluation teaching data through the lens of MFRM (Börkan, 2017).

Despite the fact that all the modelling approaches had weaknesses, the MFRM which was the strongest of all the measurement theories were found to be under-utilised. That is, the results from CMT and GT were seen as inaccurate and narrow regarding their contributions to literature in the area (Cronbach et al., 1972; Kane, 1982). GT was, however, found to have some unique features which were not present in MFRM. Based on this, there was the need for a study which combines both approach and this was exactly what this study sought to do.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

CHAPTER THREE

RESEARCH METHODS

The research methods of this present study are discussed in this section. The details of the research methods included research paradigm, research approach, design for GT, design for MFRM, study area, population, sampling procedure, data collection procedure, ethical considerations, data management, and data processing and analysis.

Research Paradigm

This study was grounded in the positivists' paradigm. Based on this, the researcher operated contingent on the belief that reality is stable and objective, and as a matter of fact can be observed and measured (Cohen, Manion, & Morrison, 2013). In this study, teachers' quality of teaching, as a variable, is characterised as stable, such that a teacher who is proficient in teaching and instruction should be rated consistently by all students who have experienced the teacher. Due to this, the variable "teaching ability" is observable, objective and can be measured. Also, the researcher was guided by the ideologies of demonstration, causal links and verification between the bits of information utilised and recognise factors that affect outcomes (Creswell, 2009). In line with this study, students appraise their lecturers by assigning values to show the extent to which the lecturer delivers particular course content. As such, the method of investigation for this study was primarily quantitative which is in line with the ideologies of the positivists.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Research Approach

This research utilised the hybrid quantitative approach. This approach to investigation combines either qualitative and quantitative methodology, two qualitative strategies, or two quantitative methodologies to obtain a better understanding of a set of data (Esgin & Senkul, 2011). This study employed two distinct approaches, which were GT and MFRM. These approaches had quantitative methodologies which were used to examine the degree of dependability of ratings provided by students as a measure of the quality of teaching. The reason for the use of this approach was to provide complimentary grounds for the use of GT and MFRM and to explore the variations in students' appraisal of teaching. That is, the MFRM was used in this study to provide detailed information to complement the results of the GT (Carroll, 2019).

Design for Generalizability Theory

Several GT designs are utilised for numerous purposes. The basic design used in this study was a two-facets partially nested random unbalanced design. A facet is a set of related conditions of measurement (Brennan, 2001b). Item, for example, was considered as a facet. Similarly, rater (i.e., student) was also a facet in this study. Item was classified as a facet because there are several items that are referred to as conditions of measurement. For this study, 13-items were used for the measurement of the teaching construct and thus, the item facet had 13 conditions of measurements. Similarly, the rater facet had over 2700 raters and so the conditions of measurement for the rater facet is over 2700. Although the object of measurement (i.e., lecturer) had several conditions of measurement, it was not considered as a facet as indicated in the GT framework (Shavelson & Webb, 1991). The sources of variations were labelled as follows:

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

lecturer (i.e., object of measurement) is symbolised as ***p***; students (i.e., rater) is symbolised as ***r***; and the item is symbolised as ***i***. Consequently, the GT design aforementioned is symbolised as ***(r:p) x i***. This design is also called a split-plot design (Cronbach et al., 1972) and has been used by researchers such as Kane, Gillmore, and Crooks (1976). A graphical and tabulated view of this design is shown in Figure 4 and Table 2 respectively.

Generalizability Theory (GT) design can be crossed (*x*), nested (*:*), or a combination of both. A design is crossed when all the conditions of measurement in a particular facet, say item facet, is observed with all the conditions of measurement of another source of measurement (e.g., raters) (Shavelson & Webb, 1991). In a study, for example, 94 students (i.e., raters) may be required to rate their classroom teacher on the quality of his/her teaching method employed. If the investigator is interested in a single facet such as rater, then the design will be person crossed with rater (*p x r*). This means that all the rater assessed the teacher's quality of teaching method used. However, if the investigator is interested in two facets, then, the design will be persons crossed with raters crossed with items (*p* x *r* x *i*).

A facet is considered nested, however, when "two or more conditions of measurement of that facet is observed with one and only one condition of measurement of another facet(s)" (Shavelson & Webb, 1991, p. 11). In this study, for instance, the rater facet was nested in the object of measurement (i.e., lecturers). That is, not all the students rated the lecturers' quality of teaching. A design can be fully nested or partially nested (Brennan, 2001b). This study employed a partially nested design because not all the facets were nested in the

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

object of measurement (i.e., lecturer). That is, each lecturer was rated on the same items even though different students rated different lecturers.

Facets in the GT framework can be random or fixed. A facet is random when the sample size is much smaller than the magnitude of the universe, and the sample is either randomly drawn or is deemed as replaceable with any other sample of the same size sampled from the universe (Brennan, 2001b). For instance, the item facet is considered random when the items used for the study is not exhaustive and that, other items can be added to perform the same function. That is, the items used in that particular study is just a sample of all the items which can function in a similar or same way. In other words, if in a GT study, there are 13 raters and other raters, who can perform the same role, can be employed to either replace or add to the existing raters, then the rater facet is random.

On the other hand, a facet is fixed when the conditions of measurement of the facet are exhausted in the universe to which the investigator would want to generalize (Shavelson & Webb, 1991). In a GT study which is conducted in Senior High School, the "core subject", for example, can be considered as a facet. If the study sample considers 4 conditions of measurement (i.e., Mathematics, Integrated Science, Social Studies, and English Language) for the core subject facet, such facet is considered as fixed because the conditions of measurement in the particular facet are exhaustive and irreplaceable, and that there is no other core subject aside these 4. All the facets in this research were treated as random.

For a GT design to be considered as balanced or unbalanced lies in whether the design has no missing information and, for any nested facet, the

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

size of the sample is unequal or constant for each level of that facet (Brennan, 2001b). This study, in particular, adopted an unbalanced design because the nested facet is unequal across the object of measurement. That is, the students who rated the lecturers were not the same from one lecturer to another. In this regard, the students are nested within the lecturers.

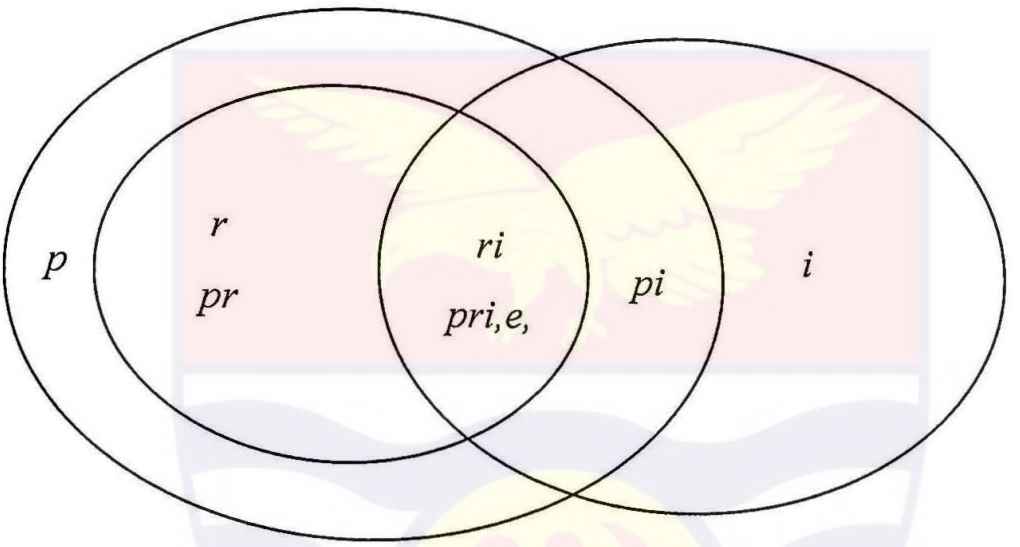


Figure 4- *Sources of Variability for* (**r: p**) x **i** *Design*

Based on the design employed, which is the two-facet partially nested random unbalanced design, the observed score for one instructor can be decomposed as follows:

$$
\begin{aligned}
X_{pir} = \ & \mu \\
& + (\mu_p - \mu) \\
& + (\mu_i - \mu) \\
& + (\mu_{pr} - \mu_p) \\
& + (\mu_{pi} - \mu_p - \mu_i + \mu) \\
& + (X_{pri} - \mu_{pi} - \mu_{pr} + \mu_p)
\end{aligned}
$$

$$\sigma^2(X_{pri}) = \sigma_p^2 + \sigma_i^2 + \sigma_{r.pr}^2 + \sigma_{pi}^2 + \sigma_{ri.pri,e}^2$$

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Table 2- *Two-Facet, Partially Nested Random Unbalanced G-study of Students' Appraisal of Lecturers*

Sources of variation	Variance component	Expected Mean Square
Person (*p*)	σ_p^2	$n_r n_i \sigma_p^2 + n_i \sigma_{r.pr}^2 + \sigma_{ri.pri,e}^2$
Item (***i***)	σ_i^2	$n_p n_r \sigma_i^2 + n_r \sigma_{pi}^2 + \sigma_{ri.pri,e}^2$
Person x item (***pi***)	σ_{pi}^2	$n_r \sigma_{pi}^2 + \sigma_{ri.pir,e}^2$
Rater: person (***r:p***)	$\sigma_{r.pr}^2$	$n_i \sigma_{r.pr}^2 + \sigma_{ri.pri,e}^2$
(Rater: person) x item (***r:p)i,e***	$\sigma_{ri.pri,e}^2$	$\sigma_{ri.pri,e}^2$

The two-facets partially nested random unbalanced design [(***r: p***) *x* ***i***] has five sources of variability, which are person (*p*), item (*i*), items crossed with persons (*pi*), raters nested in persons (*r:p*), and raters nested in persons crossed with items (*r:p*)*i,e* (see Table 2).

Design for Many-Facet Rasch Modelling

Unlike the GT, the three-facets design was employed for the MFRM framework. The facets were lecturer (i.e., object of measurement), student (i.e., rater) and item. The object of measurement, which is lecturer, was considered as a facet in the context of MFRM, although it was not considered as a facet under the GT framework. MFRM is categorised into the Rating Scale Model (RSM) and Partial Credit Modelling (PCM). The premier polytomous model is RSM which adds a threshold parameter to signify the relative difficulty of a shift from one category of the rating scale to another (Andrich, 1978; Andersen, 2005). The RSM adds a threshold parameter to represent the relative difficulty of a transition from one category of the rating scale to the next. Specifically, the category coefficient, or threshold parameter, τ_k, is the position on the latent

SAM JONAH LIBRARY
UNIVERSITY OF CAPE COAST
CAPE COAST
Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

dimension where the adjacent categories, k and k-1, are similarly likely to be observed. Stated differently, τk signifies the shift point at which the likelihood is .50 of a respondent answering in one of two adjacent categories, given that the respondent is in one of those two categories (Andrich, 1998; Bond & Fox, 2015; Linacre, 2010).

The RSM accepts the same set of threshold parameters across the rating scales. Consequently, the RSM should only be adopted when the items have the same rating scale structure; that is when the number of response categories is the same, and the relative difficulty between categories is equal across the items (Andrich, 2005; Linacre, 2006). Contrary to what happened in this research, all the 13-items were not measured using the same scale and this called for the use of PCM. As have been stipulated in literature, 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, 1982; Masters, 2010). PCM computes threshold parameters for every item independently, permitting each item to have a distinctive rating scale structure. This study adopted partial credit modelling because the instrument used in gathering the data have different response categories across the evaluation items.

$$\text{Ln}\left[\frac{p_{nik}}{p_{nik-1}}\right] = \theta_n - \delta_l - \alpha_j - \tau_k$$

where

p_{nik} = likelihood of a lecturer n obtaining a rating of k from student rater j on task l,

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

p_{nik-1} = likelihood of a lecturer n obtaining a student rating of $k-1$ from rater j on task l,

θ_n = teaching ability of lecturer n,

δ_l= difficulty of task l,

α_j = severity of student rater j,

τ_k = difficulty of getting a rating of k relative to $k-1$.

Since the MFRM employed the PCM, an adjustment needs to be done with regards to the general MFRM model. Thus, the log odds of PCM was taken into consideration. The log odds form of the PCM is given by:

$$\ln\left[\frac{p_{nik}}{p_{nik-1}}\right] = \theta_n - \delta_l - \tau_{ik}$$

Where

$p_{nik}$is the probability that a lecturer n responds with category k to item I;

p_{nik-1} is the probability that a lecturer n responds with category $k-1$ to item I;

k is a response category of a rating scale that has $m_i + 1$ categories (i.e., $k = 0, \ldots, m_i$);

τ_{ik} is the difficulty of responding with category k to item I (relative to $k-1$).

Study Area

The University of Cape Coast is one of the government universities in Ghana and is situated in Cape Coast, Central Region of Ghana. The university was established in 1962 out of a dreadful need for highly skilled and qualified human resource in education. The institution was established to offer training for graduate teachers for secondary and pre-tertiary schools in the country like technical institutions and teacher training colleges, a mission that the two prevailing government universities at the time were unfitted to accomplish.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Since then, the university has added to its roles, the training of health care specialists and doctors, and education administrators, agriculturalists, and planners. In fulfilment of its mission, the university re-engineered its degree programmes from Bachelor of Education, Bachelor of Science with Diploma in Education, and Bachelor of Arts with Diploma in Education to Bachelor of Arts /Bachelor of Science with non-education content and a Bachelor of Education, a professional qualification in Education.

The university now has courses in Bachelor of Education (B.Ed.), Bachelor of Science (B.Sc.), Bachelor of Arts (BA), Bachelor of Laws (Literally Legum Baccalaureus, LLB), Bachelor of Commerce (B.Com.), Master of Science (M.Sc.), Master of Arts (M.A.), Master of Education (M.Ed.), Master of Business Administration (M.B.A), Master of Philosophy(MPhil), Master of Commerce (M.Com.), Bachelor of Medicine/Surgery (M.B.Ch.B.), Doctor of Optometry (O.D.), and Doctorate of Philosophy (Ph.D.).

The University of Cape Coast today has five colleges which include: College of Humanities and Legal Studies (CHLS), College of Distance Education (CoDE), College of Education Studies (CES), College of Agriculture and Natural Sciences (CANS), College of Health and Allied Sciences (CoHAS) and School of Graduate Studies and Research. From an initial enrolment of 155 in 1962, the University of Cape Coast now has several students enrolled in various programmes in the regular stream, sandwich and distance education.

The University of Cape Coast was chosen for two reasons. First, a number of studies have been conducted in an attempt to understand the dynamics in student ratings of instructors and their courses in UCC (Gyimah et al., 2016; Kwarteng, Doku, Matta, & Doh-fia, 2014). These previous studies

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

revealed inconsistent evaluation results over the period of time and that lecturers raised concerns about student responses as well as and (2) the researcher has experienced the evaluation system in the study setting which serve as a motivation to conduct such a study.

Population

The study covered regular students in UCC. The population was targeted to regular students which were 24,726 under four colleges in the university for the 2019/2020 academic year. Regular students were the focus of the study because it was the only available data as at the time of the study. This comprised both regular undergraduate and postgraduate (course-work level) students clustered under these colleges: 4,846 students from College of Agriculture and Natural Sciences (CANS), 9,717 from College of Humanities and Legal Studies (CHLS), 6,429 from College of Education Studies (CES), and 3,734 from College of Health and Allied Sciences (CoHAS). The College of Distance Education (CoDE) was not involved because the college does not have regular students.

In terms of responding to the online evaluation form, 152,658 responses were expected from the 24,726 students based on 1,673 courses registered that semester. This was because it was expected that students respond to the evaluation form based on the number of courses registered. Assuming a student registered for 5 courses for the second semester of the 2019/2020 academic year, such student was expected to respond to the evaluation form 5 times, each response for each course.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Sampling Procedure

The study used secondary data. These data are information already obtained from the students for decision-making purposes. All cases and data points were involved in this study. Thus, the sample size was 24,726 regular students corresponding to 152,658 expected responses (based on the number courses taken by the students) from 1,673 courses. It should be noted that students took varying number of courses. Out of the 152,658 expected responses, only 73,906 responses were received from the students by DAPQA-UCC. That is, not all students responded to the online evaluation form. This number of students involved in the study through a census approach. Table 3 presents the summary of courses/instructors evaluated for the 2019/2020 second semester academic year.

Table 3- *Details of Courses Assessed for 2019/2020 Second Semester*

Level of study	No. of courses	Percentage
Level 100	196	11.7
Level 200	305	18.2
Level 300	416	24.9
Level 400	384	23.0
Level 500	8	.5
Level 600	5	.3
Level 700	4	.2
Level 800	318	19.0
Level 900	37	2.2
Total	1673	100.0

Source: DAPQA, UCC (2021)

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

A total of 1,673 courses were evaluated by the regular students for the second semester of the 2019/2020 academic year. Out of this, the majority of the courses evaluated were level 300 courses (n=416, 24.9%). The courses evaluated by the levels 200 (n=305, 18.2), 400 (n=416, 24.9%) and 800 (n=318, 19%) students were also quite high. Few of the courses evaluated were by levels 500 to 700 students.

It must be stated that the data obtained for this research were screened and cleaned based on four criteria which include: (1) courses that had only one student responding were deleted from the final data set since there will be no variance; (2) core courses were deleted from the analysis because such courses were merged in the data provided by DAPQA-UCC; (3) duplicated courses were deleted because this could confound the findings of the study; (4) specific courses which had inconsistencies in terms of the responses were deleted. After the implementation of these criteria, 2,553 students (within 145 courses) remained in the data set for final analysis.

Data Collection Instrument

The data for this study was collection using the instrument developed by DAQPA-UCC. The research solely used this data from the university which were later on analysed. However, the instrument that was used for the evaluation survey by the university was described. This is to give an idea of the data gathered by UCC.

The evaluation form used by UCC has 25-items comprising 22 closed-ended and 3 open-ended items (see Appendix A). 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-

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

items). In this study, the course outline dimension was removed from the data before use 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 so factual such that is variability in this item 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, items cease to be a facet in the course outline dimension.

Also, two other sections (i.e., attendance and assessment) on the appraisal evaluation form were discarded and data on these areas were not included in this research. These dimensions comprised items that were objective and did not require any subjectivity in the scoring resulting in approximately zero variances. An example of an item under the assessment section was to find out the number of assessments given by the lecturer. Options to this item were (a) 0, (b) 1, and (c) 2 or more. Answering this statement of this nature would require counting the number of assessments provided by the lecturer, and not engaging in any subjective rating. Similarly, a statement captured under the attendance section required students to indicate how many times they met the lecturer. Responding to this item will also require just recalling the number of times the lecturer has met the class.

The 3 open-ended items required students to write the lecturer's strength(s), weakness(es), and any other suggestions. The responses from the open-ended were also not included in this research because the responses provided were not quantitative. It becomes practically impossible to quantify these qualitative responses due to the large volume of the data. In addition to this, the majority of the students do not usually respond to the three open-ended

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

questions. Only two sections of the evaluation form were involved in this study, that is, the mode of delivery and course contents. The items for these two sub-scales were 13 in all.

The validity and reliability of the evaluation form are not well understood in its current form. That is, there are relatively few documented materials on the validity and reliability of the items used in the evaluation process. Institutional documents which the researcher have come across only mentions that the evaluation form is well developed and validated (UCC, 2012). However, a study conducted by Kwarteng and Doh-fia (2014) examined the psychometric properties of the instrument and revealed a Cronbach Alpha reliability estimate of .94. The authors also conducted a factor analysis and confirmed most of the items were sufficient, except for a few ones.

Data Collection Procedure

Ethical clearance (see Appendix C) was obtained from the Institutional Review Board (IRB) of UCC after the proposal had been approved. An introductory letter was also obtained from the Department of Education and Psychology, UCC (see Appendix B). A letter was also obtained from one of the supervisors (see Appendix E), together with which the researcher drafted another letter (see Appendix D) to officially seek permission to access and use the data. Copies of these letters were sent to DAPQA-UCC (specifically addressed to the Acting Director) to formally ask for permission to access and use the data. This was done after there has been an initial contact with the Director of DAPQA-UCC to pre-inform the directorate and discuss with them what the study sought to achieve and the need for the study (Creswell, 2012).

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Ethical Consideration and Data Management

The study took into consideration ethical issues such as the IRB clearance, confidentiality, anonymity, and data de-identification. The researcher applied for ethical clearance from the IRB 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' identity were removed from the data to make all students anonymous. In addition, the identity of the lecturers/courses evaluated was hidden and they were given pseudonyms. This was done through a process called data de-identification. For example, the lecturers were named AA, AB, AC, AD, AE, etc.

For informed consent, this was provided by DAPQA. This was because students during the administration phase permitted the university to use the data for managerial and administrative purposes as well as for future research use (Jol & Stommel, 2016). Based on this, informed consent has been provided in a retrospective which also empowers the university to give the data to any party for the purposes stated above.

The data were acquired in a soft copy form and were validated manually by taking the time to go through the data. The validation process also included checking for the accurateness of data and ensuring that the ratings provided falls within the rating scale. The data were stored on a laptop with a password that was only known to the researcher. The data were processed and presented in a manner that there had no traces of what a particular respondent provided.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Data Processing and Analysis

The data were screened and cleaned. The analyses were conducted based on the modelling approaches used. Before the major analysis, the data were explored to have a fair idea of the nature of the data.

GT analysis

Generalizability Theory (GT) offers a platform for segregating the aggregate variance in a set of scores into distinct, unrelated components which are respectively linked with a different source of variability (Brennan, 2001b). It must be indicated that GT analysis is in two phases: Generalizability study (G-study) and Decision study (D-study). The first phase, which is G-study, computed for estimates of the variance components related to the universe of admissive observation for the respective facets. The variance values were interpreted using a 100% cut-off through the total variance. A G-study aims to "anticipate the numerous uses of measurement and to offer as much information as probable about the sources of variances in the measurement" (Shavelson & Webb, 1991, p. 12).

With regard to this study, where lecturers who handled the various courses are rated by students using several evaluation items, GT partitions the total variability in the scores into the following variance components: (1) systematic inconsistencies between separate lecturers, (2) raters variability (interrater variations), and (3) variability between items (Brennan, 2001a). These variance components are known as the main effects which are similar to the main effect in the analysis of variance. GT, in addition to this, estimates for variance components for interaction between or among the main effects were obtained. Up to this point, the relative magnitude of the estimated variance

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

components was estimated to find out which source(s) of variance is/are producing measurement error and which source of variance(s) needed to be looked at to decrease unnecessary variations in scores (Sudweeks et al., 2005).

The second phase of GT analysis, known as a decision study (D-study), information from the variance components from the G-study was used for designing measurement procedure which reduces measurement errors for a specific purpose. The investigator specified the universe of generalisation (i.e., the universe which the decision-maker or investigator desires to generalise depending on the obtained information in a G-study) (Brennan, 2001b, 2011; Marcoulides, 2000; Shavelson & Webb, 1991).

Two main forms of decision-making conditions were highlighted. In situations where ratings were employed for decision making depending on the relative standing of the examinees (i.e., lecturers in the case of this study), the decision was known as relative decisions. Thus, the relative standing of every individual is not only determined by their ratings but also the ratings of other individuals within the normative group (Brennan, 2001b). Where decisions were made based on whether individuals rating exceed or equal some criterion or least tolerable standard of ratings on the construct, the decision type was known as absolute decision (Cronbach et al., 1972). In this case, the conclusions about every single object of measurement were independently reached regarding how other persons performed. The discrepancy between these two main forms of decisions was significant because the description of what constitutes errors of measurement was dependent on which kind of conclusions drawn (Brennan, 2001a; Ebel, 1951). In the D-study analysis, the investigator took the following steps:

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

1. The universe of generalization was first described--the breadth and number of facets that the investigator is eager to generalise over;
2. The projected interpretation of the measurement was stated -- absolute or relative; this projected interpretation describes the measurement error and in so doing recognises the sources of variation of prime concern; and
3. information from the G-study (regarding the degree of the sources of measurement errors) was used to estimate the efficiency of other designs for reducing error and improving reliability.

The GT analyses highlighted key estimates upon which conclusions were made. These include:

1. The absolute error variance

All variance components excluding the variance resulting from the facet of concern, known as the object of measurement in GT and referring to lecturers in the context of this study, add up to the error of measurement when ratings are utilised for making absolute decisions. Therefore, the absolute error variance is the summation of all variance components except for the variance component for the object of measurement, which is omitted since it is not deemed as error variance. Rather, it signifies systematic variance in the mean scores of the different lecturers (averaged across all raters, and all items) and is equivalent to true-score variance in CMT (Brennan, 2001a). For the two-facet partially nested random unbalanced design, the absolute error variance is given by the formula:

$$\sigma^2_{(\Delta)} = \sigma^2_i + \sigma^2_{pi} + \sigma^2_{r.pr} + \sigma^2_{ri.pri,e}$$

where,

$\sigma^2_{(\Delta)}$ is the symbol for absolute error variance

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

σ_i^2 is variance component resulting from items;

σ_{pi}^2 is variance component resulting from persons crossed with items;

$\sigma_{r.pr}^2$ is variance component resulting from raters nested in person,

$\sigma_{ri.pri,e}^2$ is variance component resulting from raters nested in persons crossed with items plus other unknown errors

2. The relative error variance

Only those variance components which represent the interactions with the object of measurement (lecturer in this study) contribute to this error of measurement and associated with relative decisions. The relative error variance, therefore, is the summation of all variance components within a specified model which demonstrate an interaction between the object of measurement and any of the facets. The square root of this estimate is equivalent to the standard error of measurement in CMT (Brennan & Johnson, 1995).

$$\sigma_{(\delta)}^2 = \sigma_{pi}^2 + \sigma_{r.pr}^2 + \sigma_{ri.pri,e}^2$$

where,

$\sigma^2(\delta)$ is a symbol for relative error variance

σ_{pi}^2 is variance component resulting from persons crossed with items;

$\sigma_{r.pr}^2$ is variance component resulting from raters nested in persons;

$\sigma_{ri.pri,e}^2$ is variance component resulting from raters nested in persons crossed with items plus other unknown errors

3. The generalizability (*g*) coefficient

The *g*-coefficient is used for making relative decisions and, thus, associated with the relative error variance. This is a reliability-like estimate which ranges from 0 to 1.0. The higher the reliability estimate, the better the reliability of the data. It is estimated by dividing the systematic variance in the

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

average ratings of the object of measurement (i.e., lecturer) by summing the relative error variance and the systematic variance (Brennan, 2001a). The computation formula for estimating the *g*-coefficient is:

$$g = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{Rel}^2}$$

4. The dependability or phi-coefficient

The dependability coefficient is used for making absolute decisions and thus, associated with the absolute error variance. This is a reliability-like estimate which ranges from 0 to 1.0. The higher the reliability estimate, the better the reliability of the data. It is estimated by dividing the systematic variance in the average ratings of the object of measurement (i.e., lecturer) by summing the absolute error variance and the systematic variance (Brennan, 2001a). The computation formula for estimating the phi-coefficient is:

$$\Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{Abs}^2}$$

It is important to note that the two co-efficients (i.e., phi and g) are exceptional cases of intra-class correlation. These two coefficients have a related structure that is equivalent to the structure of the reliability estimate in CMT (Crocker & Algina, 1986). The discrepancy between the two estimates depends on the description of what constitutes an error for the kind of decision to be made. It is important to highlight that both phi and g coefficients (in this study) were evaluated using the approach of Creswell (2012) who indicated that researchers investigating validity issues through some correlation means should use a cut-off of .86 and above as adequate to attain a high construct validity.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Creswell further indicated that coefficients from .66 to .85 can be considered as good.

The choice of software for GT analyses

The statistical software used for GT analyses started with Brennan developing a computer software programme in 1983 which was known as GENOVA (Generalised Analyses of Variance). Brennan designed one for univariate analysis (called the urGENOVA software), and another for multivariate analysis (known as the mGENOVA software) (Brennan, 2001b) The mGENOVA programme was developed to handle multivariate variables and the urGENOVA package used for the estimate of the variance components of both unbalanced and balanced random designs (Brennan, 2001c).

Though the programmes designed by Brennan are regularly utilised by scholars or researchers working on GT, the intricacy of the packages instigated challenges, particularly for researchers using for the first time. The absence of a manageable computer programme constrained the prevalent utilisation of GT for several decades (Cardinet, Johnson, & Pini, 2010; Güler, Uyanik, & Teker, 2012). In searching for a solution to this challenge, Duquesne from Belgium, in the early 1980s, designed a computer programme known as ETUDGEN. Although this programme satisfied the basic academic desires of scientists in the area of measurement and beyond, it had a limitation in terms of conducting complex analysis (Cardinet et al., 2010).

At that time, Mushquash and O'Connor (2006) had provided their syntaxes which they developed in their 2006 article to run GT analyses for computer software like MATLAB, SAS, and SPSS. Following their instructions, it became practically easy to conduct GT analyses with these

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

computer programmes which were already commonly used by several scientists. This approach to GT analysis offered a platform where both G and D studies were conducted and graphs for the phi coefficient, G coefficient absolute error variances, and relative error variances can be produced. There were two distinct syntax groups for GT analysis using SPSS: G1 and G2. The G1 syntax assisted researchers to carry out analyses for single and two-facet designs. The G2 syntax, on the other hand, analysed three-facets designs although this involved complex procedures. Although two-facet nested and crossed situations were easily analysed with the G1 syntax on MATLAB, SAS, and SPSS packages, more than two-facets that were crossed were only possible using SAS and SPSS programmes through several complex procedures in the G2 syntax.

When using SPSS, the estimates from the D studies were carried out distinctly for absolute and relative error variances, and the phi and G coefficients were also provided using graphs. The challenge with the use of SPSS was that it was only possible to use random designs. In addition to this, manual computations needed to be done for fixed facet designs in SPSS. This weakness was addressed by the introduction of EduG. However, both EduG and SPSS packages were found to have limited ability in conducting GT analysis for nesting designs with unbalanced data (Teker, Güler, & Uyanik, 2015).

The EduG was designed in 2006 by Cardinet which was found to have user-friendly and a versatile platform for performing GT analyses (Cardinet et al., 2010). The number of facets involved in the GT analyses using EduG can be increased as large as possible. That is, there is practically no limit to the facet number during the analyses. Nevertheless, the non-existence of graphs in EduG has been cited as a challenge in the measurement literature (Güler et al., 2012).

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

According to Musquash and O'Connor (2006), EduG package should be used for all situations where the data is balanced because of the several structures offered by EduG in the manipulation of data.

As have been earlier mentioned, EduG cannot handle unbalanced designs. This limitation was resolved by introducing the G-String package. Even though G-String could handle all unbalanced designs, there was some weakness in handling crossed, mixed-model balanced designs. In addition to this, G-String also provides lower reliability estimation for absolute decisions. LISREL programme also allows for GT analysis, however, it only yields outputs for only G-study, without presenting for D study outcome (Teker et al., 2015).

R software is the only software that provides a rich set of resources for applying GT. The R programme is free and modern, as well as provide high-quality visualizations for reporting G-theory results (Hucbner & Lucht, 2019). The use of R makes GT analyses to the extent that any analyses can be done. Hence, the R software was used for the GT analyses for this study.

MFRM analysis

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 raters' severity, variances in the relative difficulty of the task (Lunz, Wright, & Linacre, 1990). It must be indicated that, in this study, the person facet was positively arranged (i.e., high on the logit scale denotes high ability), raters facet was negatively arranged (high on the logit scale means more severe rating and vice versa), and item facet was also negatively ordered (i.e., high on

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

the logit scale represents more difficult items). A software called FACET was used for the MFRM analysis. The analysis of MFRM provided the following indexes and assessment criteria:

1. The Infit and Outfit mean square (MnSq) estimate

The Infit and Outfit MnSq estimate denote the difference between observed and the expected model-driven responses and flag unexpectedness in the data set (Linacre & Wright, 2002). The value of these estimates ranges from zero to infinity. In the event of perfect correspondence, the value of these estimates become 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 rate unpredictably and inconsistently. 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, Wu and Adams (2013) argued that the infit and outfit statistics are sensitive to sample and, as results, caution should be taken when deciding on the upper and lower control limit for these estimates. They, therefore, conducted several simulations and designed the appropriate lower and upper control limit 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.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

The formula for the outfit mean square is given by

$$MS_{U(J)} = \frac{\sum_{n-1}^{N} \sum_{i=1}^{I} z_{nij}^2}{N \cdot I}$$

The formula for the infit mean square is given by

$$MS_{W(J)} = \frac{\sum_{n-1}^{N} \sum_{i=1}^{I} w_{nij}\, z_{nij}^2}{\sum_{n-1}^{N} \sum_{i=1}^{I} w_{nij}}$$

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 a person (i.e., lecturer) and item facet, a low G value is preferred for the rater (i.e. student) facet. If the separation index (i.e., rater separation index) is calculated for raters, for instance, *g* index is given by:

$$G_R = \frac{SD_{t(R)}^2}{RMSE_R}$$

where,

$RMSE_R$ is root-mean-square measurement error associated with rater severity measures

$SD_{t(R)}^2$ is the true score variance of rater severity measures

3. The reliability of Separation Index (R)

The reliability of the separation index is symbolised by R. This reliability of separation index illustrates how reproducibly different the

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

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). Just the same way G value is interpreted, a high R-value is preferred for a person (i.e., lecturer) and item facet, whereas a low G value is preferred for the rater (i.e., student) facet. The reliability of separation index (R) for rater, for example, is given by:

$$R_R = \frac{SD^2_{t(R)}}{SD^2_{O(R)}}$$

where,

$SD^2_{t(R)}$ is denoted by true score variance for the rater facet

$SD^2_{O(R)}$ is denoted by observed score variance for the rater facet

4. Fixed chi-square statistics (all-same statistics)

The fixed chi-square statistics 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. MFRM Interaction/bias analysis

In MFRM, an interaction between any two facets is termed as bias analysis. The interaction between facets was assessed by using standard scores (z-score). In an instance where an absolute estimate of *t*-score is larger than 2.0, then, there is evidence or indication of a statistically significant interaction between some specified facets.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

6. 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 low relatively low frequencies and percentages show non-functional rating scale and such categories need to be revised or combined with other scale categories.

Also, the category average measure is another indicator that can be examined to understand scale functioning. In this case, the average logit measure of a particular scale category should be either 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.

Table 4- *Summary of Analytical Procedure*

No.	Research Questions	Analysis Remark
1.	What are the sources of measurement errors (raters and items) in students' appraisal of courses and teaching?	G-study was conducted. Variance components for sources of variation were calculated: (1) person, (2) items nested in persons; (3) raters nested in persons; (4) raters crossed with items and all nested in persons. Percentage variances and mean squares of the potential sources of variances were also presented.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

2.	What is the dependability of students' ratings of the quality of teaching and course contents?	The generalizability- and phi-coefficients were presented and interpreted to explain the level of dependability. Much emphasis was premium was placed on generalizability coefficients. These were presented alongside their measurement errors.
3.	How many raters and items are needed to obtain an optimum level of dependability in students' appraisal of courses and teaching?	D-study was conducted. Several optimisations of the rater and item facets were carried out.
4.	How does non-response influence the validity of students' evaluation of courses and teaching?	Optimisation on the number of raters holding all other facets constant was carried out. Charts were also used for further presentation.
5.	What are the behaviours of students (halo effect, and severity effect) in the appraisal of courses and teaching?	Rater severity was examined. Separation index and separation reliability for raters were computed. Infit and Outif indices were also used.
6.	How effective are evaluation items in the measurement of the quality of course content and teaching?	Item difficulty was examined. Separation index and separation reliability for items were computed. Infit and Outif indices were used. Quality control indicators were also used to judge the quality of the items.
7.	How fair are lecturers rated in terms of the quality of lecturers' teaching?	The fairness estimate for instructors was examined against their observed rating. The separation index and separation reliability for instructors were computed. Infit and Outif indices were also used.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

8.	What are the biases (if any) in students' appraisal of courses and lecturers' teaching?	Exploratory interaction bias analysis was conducted. Bias size parameter together with their *t* and *p*-values were examined to determine whether an interaction was present.

RQ- Research Question

Chapter Summary

The study was grounded in the positivists' paradigm using the hybrid quantitative approach. Two main quantitative designs were used: a two-facets partially nested random unbalanced design and the three-facets design with a partial credit model. The study covered students in UCC. The population was targeted to regular students which were 24,726 under four colleges in the university for the 2019/2020 academic year. The study used secondary data on students obtained from the Directorate of Academic Planning and Quality Assurance, UCC (DAPQA-UCC). All cases and data points were involved in this study. Thus, the sample size was 24,726 regular students. This number of students involved in the study through a census approach.

Ethical clearance was obtained from the Ethical Review Board of UCC after the proposal has been approved. An introductory letter was also be taken from the Head, Department of Education and Psychology. A letter was also obtained from one of the supervisors, together with which the researcher drafted another letter to officially seek permission to access and use the data. Copies of these letters were sent to DAPQA-UCC (specifically addressed to the Acting Director) to formally ask for permission to access and use the data. This was done after there has been an initial contact with the Director of DAPQA-UCC to pre-inform the directorate and discuss with them what the study sought to achieve and the need for the study (Creswell, 2012). Other ethical

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

considerations taken into account were confidentiality, anonymity, and data de-identification. The data were analysed by conducting univariate GT and partial credit MFRM analyses.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

CHAPTER FOUR

RESULTS AND DISCUSSION

This chapter highlights how the data were analysed and presented. As already indicated, the data obtained were responses from students on the appraisal of courses and teaching. This chapter first presents the results for each research question, followed by the discussion. The results section was presented in two separate sections based on the two theories; GT and MFRM. The analyses on research questions 1-4 were conducted using GT whereas MFRM was used to analyse data to address research questions 5-8.

Research Question 1

What are the sources of measurement errors (raters and items) in students' appraisal of courses and teaching?

This research question examined the sources of measurement errors in students' appraisal of courses and teaching. Based on the GT design used, five sources of measurement errors were identified, which were person (p), item (i), person x item (pi), rater: person ($r{:}p$), and (rater: person) x item ($r{:}p)i,e$. Their variance components and percentages were further calculated to understand the contribution of each of the sources of variability to the ratings. Separate analyses were provided for appraisal of courses and evaluation of teaching quality. The details of the analysis are shown in Table 5.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Table 5- *Sources of Variability and their Variance Components for Course and Teaching Appraisal*

Sources of variation	df	SS	MS	Variance	Variance Percentage
Course Appraisal					
person (*p*)	144	787	5.463	.041	2.7
item (*i*)	2	6	3.189	.001	0
person x item (*pi*)	288	514	1.784	.044	2.9
rater: person (*r:p*)	2556	6153	2.407	.488	32.3
(rater: person) x item (*r:p*)*i,e*	5112	4791	.939	.939	62.1
Teaching Evaluation					
person (*p*)	144	4675	32.47	.150	9.20
item (*i*)	9	300	33.28	.011	.70
person x item (*pi*)	1296	3019	2.33	.063	3.80
rater: person (*r:p*)	2556	9374	1.67	.250	15.2
(rater: person) x item (*r:p*)*i,e*	23004	26842	1.17	1.167	71.1

df- degrees of freedom; MS- mean square; SS- sum of squares

The results, as shown, in Table 5 revealed that the residual [(rater: person) x item (*r:p*)*i,e*] had the largest variance (σ^2=.939, 62.1%) contribution to the students' appraisal of courses. This suggested that there is a systematic interaction of students in each class and items, coupled with other systematic and unsystematic errors which were not explored in this study. The second-largest sources of variability to students' appraisal of courses was the rater nested person (*r:p*) facet (σ^2=.488, 32.3%) indicating that students systematically differed on how they evaluated the same course. This was

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

followed by person crossed with item (*pi*) (σ^2=.044, 2.9%), and persons (*p*) (which is not considered as a facet) (σ^2=.041, 2.7%). Items had the least variance contribution to the appraisal of courses (σ^2=.00).

For the evaluation of teaching, the largest source of rating variation was due to the residual [(rater: person) x item (*r:p)i,e*] with a variance of 1.167 and a variance percentage of 71.1 (σ^2=.1.167, 71.1%). This suggests that the variability of students' rating of lecturers is influenced by the systematic interaction of raters (nested in classes) by items, as well as other systematic and random factors which were not explored in this study. Relatively, raters nested in persons (*r:p*) had the second-largest variance contribution to students' appraisal of lecturers with a variance of .250 and a corresponding percentage of 15.2. This indicates that raters systematically were different in how they rated the same lecturer. Item (σ^2=.011, .7%). had the least contribution to the variability in students' rating of lecturer signifying that there was a consistent use of the items among students from a single class.

Research Question 2

What is the extent of dependability of students' ratings of the quality of teaching and course contents?

This research question assessed the extent of dependability of students' appraisal of courses and teaching. This was evaluated using the index of dependability (Φ) and generalizability index (*g*) with their respective relative and absolute errors. Due to the nature of students' appraisal of courses and teaching, more emphasis was placed on the generalizability index. This is because of the utilisation of class means where students in a particular class need to be homogeneous in their responses since they aim to evaluate the same

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

trait of the instructor or course. This notwithstanding, understanding individual students' rating behaviours was also deemed relevant to achieve homogeneity in the appraisal process. Before the estimation of the coefficients (both the index of dependability (Φ) and generalizability index (g)), the average number of raters per class were computed based on the responses received from the students through the survey conducted (by DAPQA-UCC). This is obtained by dividing the total number of responses (N=73906) by the number of courses (n=1673) which yielded a mean of 44.17. The details of the results are shown in Table 6.

Table 6- *Reliability Estimate of Student Appraisal of Courses and Teaching*

Indicators	Course Appraisal	Teaching Appraisal
No. of items	3	10
No. of raters (Average response)	44	44
Relative Error $[\sigma^2_{(\delta)}]$	.033	.034
Absolute Error $[\sigma^2_{(\Delta)}]$	.033	.034
Index of dependability (Φ)	.554	.802
Generalizability index (g)	.556	.806

Source: Field Secondary Data (2021)

As presented in Table 6, it is instructive to highlight that the number of items for the evaluation of course content and teaching quality is 3 and 10 respectively. The results, as shown in Table 6, established some degree of validity of the data provided by students regarding the appraisal of courses and teaching. Generalizability coefficients of .556 and .806 were achieved for students' appraisal of courses and evaluation of teaching respectively. While both coefficients were good, they were not sufficient enough to establish a high

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

validity of the measurement of quality of teaching and course contents based on the recommendation of Creswell (2012) (g>.84) (for both coefficients). Nonetheless, responses provided by the students on the teaching evaluation provided by students appeared more valid than the course appraisal data.

Research Question 3

How many raters and items are needed to obtain an optimum level of dependability in students' appraisal of courses and teaching?

This research question aimed at identifying the number of raters (students) and evaluation items needed to obtain an adequate level of dependability in students' evaluation of courses and teaching. To achieve this, a D-study was conducted separately for course appraisal and teaching evaluation. For each of these, optimisation was carried out for large class size as well as small class size. This was done to understand the dynamics of student ratings in both small and large classes since such a situation exists in UCC.

Optimisation for course appraisal data

Optimisation was conducted using 7 options with each having a different number of item and rater combination. The result is presented in Table 7.

Table 7- *Optimisation for Course Appraisal for Large Class Size*

Options	1	2	3	4	5	6	7
No. of items (n_i)	5	7	9	11	13	15	18
No. of raters (n_r)	20	30	40	50	60	70	80
Relative Error $[\sigma^2_{(\delta)}]$	.043	.027	.020	.015	.013	.011	.009
Absolute Error $[\sigma^2_{(\Delta)}]$	.043	.027	.020	.015	.013	.011	.009
Index of dependability (Φ)	.490	.602	.675	.726	.763	.791	.817
Generalizability index (g)	.491	.603	.676	.726	.764	.791	.817

Source: Field Secondary Data (2021)

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

The result, as presented in Table 7, showed that with 5-items on course appraisal and 20 raters (students), a generalizability index of .491 would be achieved. This index increased to .603 when there were 7 evaluation items and 30 raters. With 9-items and 40 raters, a generalizability index of .676 was achieved. A closer look at options 6 (15-items and 70-raters and 7 (18-items and 80 raters) showed generalizability indices of .791 and .817. In effect, it appears that having 18 evaluation items and 80 raters produced the largest generalizability index for course appraisal (g=.817, Φ =.817). This indicates that option 7 produced better results.

Table 8 presents the result for optimisation for small classes.

Table 8- Optimisation for Course Appraisal Data for Small Class Size

	n_i=10		n_i=20		n_i=30	
n_r=...	Φ	g	Φ	g	Φ	g
1	.065	.065	.065	.071	.073	.073
3	.171	.171	.171	.185	.190	.190
5	.253	.253	.253	.273	.280	.280
7	.319	.319	.319	.343	.351	.351
9	.373	.373	.372	.400	.409	.409
11	.417	.417	.417	.447	.457	.457
13	.455	.455	.455	.486	.497	.497
15	.487	.487	.487	.520	.532	.532
17	.515	.515	.515	.549	.561	.561
19	.539	.539	.539	.575	.587	.587
21	.561	.561	.561	.597	.610	.610
23	.580	.580	.580	.617	.630	.630
25	.597	.597	.597	.635	.648	.648

Source: Field Secondary Data (2021)

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

The options displayed in Table 8 is more useful and practical for small class sizes. For this optimisation, the number of raters was fixed from 1 to 25. Three options were used for the number of course evaluation items. For each of the item-raters combination, the generalizability coefficient and index of dependability were presented.

The results displayed in Table 8 indicated that, with 10 appraisal items, the generalizability coefficients ranged from .065 (for only one rater) to .597 (for 25 raters). With course evaluation items increased to 20, generalizability coefficients ranging from .071 (for only one rater) to .635 (for 25 raters) were produced. Similarly, the optimisation produced generalizability indices ranging from .073 (for a single rater) to .648 (for 25 raters) when course evaluation items were increased to 30. The 20-item option was explored with the number of raters ranging from 1-50 (see Figure 5).

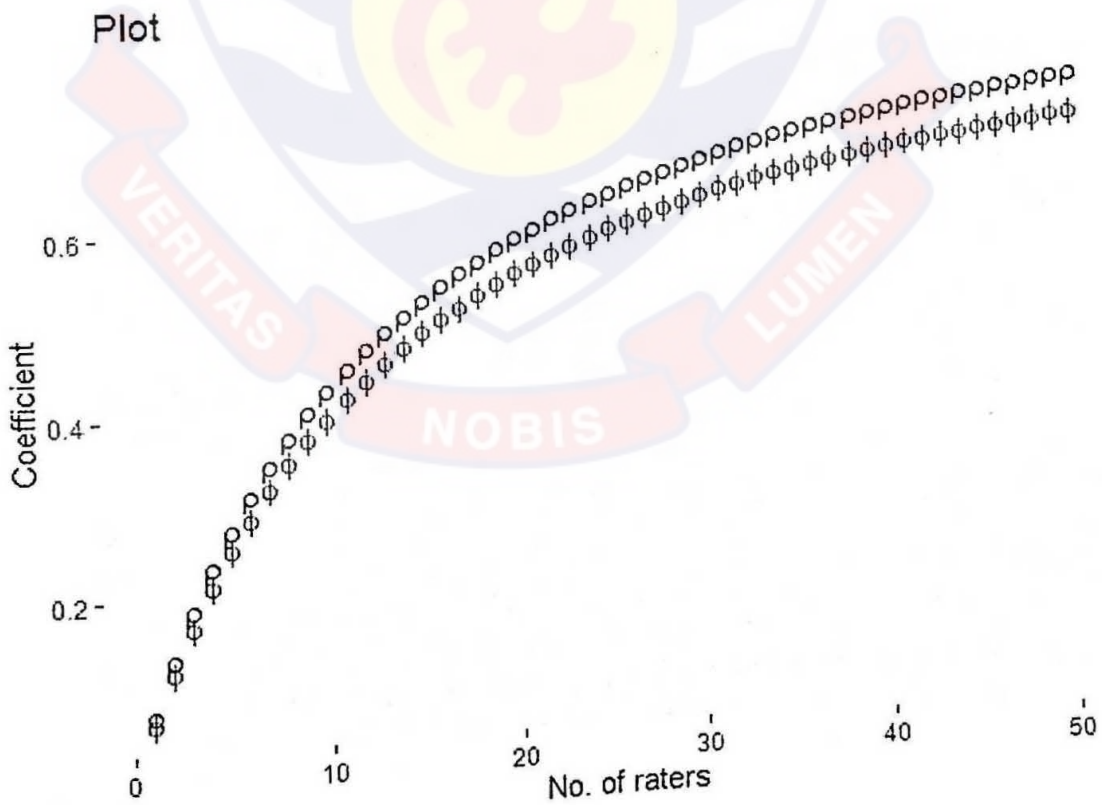


Figure 5- *Optimisation (Course) Graph with 20-items*

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Option 2 (n_i=20) appeared more practical and useful since the generalizability coefficients produced were not too different from option 3 (n_i=30). The graph showed increasing indices (both index of dependability and generalizability index) as the number of raters increase (Figure 5).

Optimisation for evaluation of teaching data

D-study was also carried out to provide practical suggestions for improving the extent of dependability of data provided by students regarding their evaluation of teaching. The details of the results are presented in Tables 9 and 10, as well as Figure 6. In Table 9, the optimisation of teaching evaluation is presented with a focus on instances where there is a large class size.

Table 9- *Optimisation for Teaching Evaluation Data for Large Class Size*

Options	1	2	3	4	5	6	7
No. of items (n_i)	12	14	16	18	20	22	24
No. of raters (n_r)	20	30	40	50	60	70	80
Relative Error $[\sigma^2_{(\delta)}]$	.023	.016	.013	.010	.009	.008	.007
Absolute Error $[\sigma^2_{(\Delta)}]$	.023	.016	.013	.010	.009	.008	.007
Index of dependability (Φ)	.865	.902	.922	.935	.945	.951	.957
Generalizability index (g)	.870	.906	.926	.939	.948	.954	.960

Source: Field Secondary Data (2021)

Table 9 displays the dependability and generalizability indices for seven optimisation options. The results (in Table 9) showed that option 7 (with 24 teaching evaluation items and 80 raters) produced the highest generalizability coefficient of .96. It is instructive to note that options 2 (14-items and 30 raters),

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

3 (16-items and 40 raters), 4 (18-items and 50 raters), 5 (20-items and 60 raters), and 6 (22-items and 70 raters) produced a high generalizability coefficient (>.90). Although options 2-7 yielded very high generalizability indices, option 1 (with 12 teaching appraisal items and 20 raters) also produced a relatively high generalizability index (g=.870) which is also adequate for the measurement of teaching quality based on Creswell's (2012) recommendation. Based on practical grounds, option 1 is deemed appropriate for decision making.

Further, optimisation was conducted to cater for small class sizes and the results are displayed in Table 10.

Table 10- *Optimisation for Teaching Evaluation with Small Class Size*

n_r=...	n_i=12		n_i=15		n_i=20	
	Φ	g	Φ	g	Φ	g
1	.287	.288	.312	.312	.325	.326
3	.537	.539	.569	.570	.586	.587
5	.651	.654	.681	.683	.697	.699
7	.716	.720	.744	.747	.759	.761
9	.758	.762	.785	.788	.799	.801
11	.787	.792	.813	.816	.826	.828
13	.809	.814	.833	.837	.846	.849
15	.826	.831	.849	.853	.861	.864
17	.839	.844	.862	.865	.873	.876
19	.849	.855	.872	.875	.883	.886
21	.858	.864	.880	.884	.891	.894
23	.866	.871	.887	.891	.898	.901
25	.872	.878	.893	.897	.904	.907

Source: Field Secondary Data (2021)

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

As presented in Table 10, optimisation with 12 teaching evaluation items yielded generalizability coefficients ranging from .288 (for one rater) to .878 (for 25 raters). With 15 teaching evaluation items, a generalizability index ranging .312 (for one rater) to .893 (for 25 raters). Generalizability coefficients between .325 (for one rater) and .907 (for 25 raters) were produced with 20 evaluation items. It appears that the coefficients were not so different from each other for cases where the items were 12, 15 and 20. Based on this, 12 evaluation items (with raters between 12-25) may be deemed useful in producing an optimum reliability index for the measurement of teaching quality. This is shown in Figure 6.

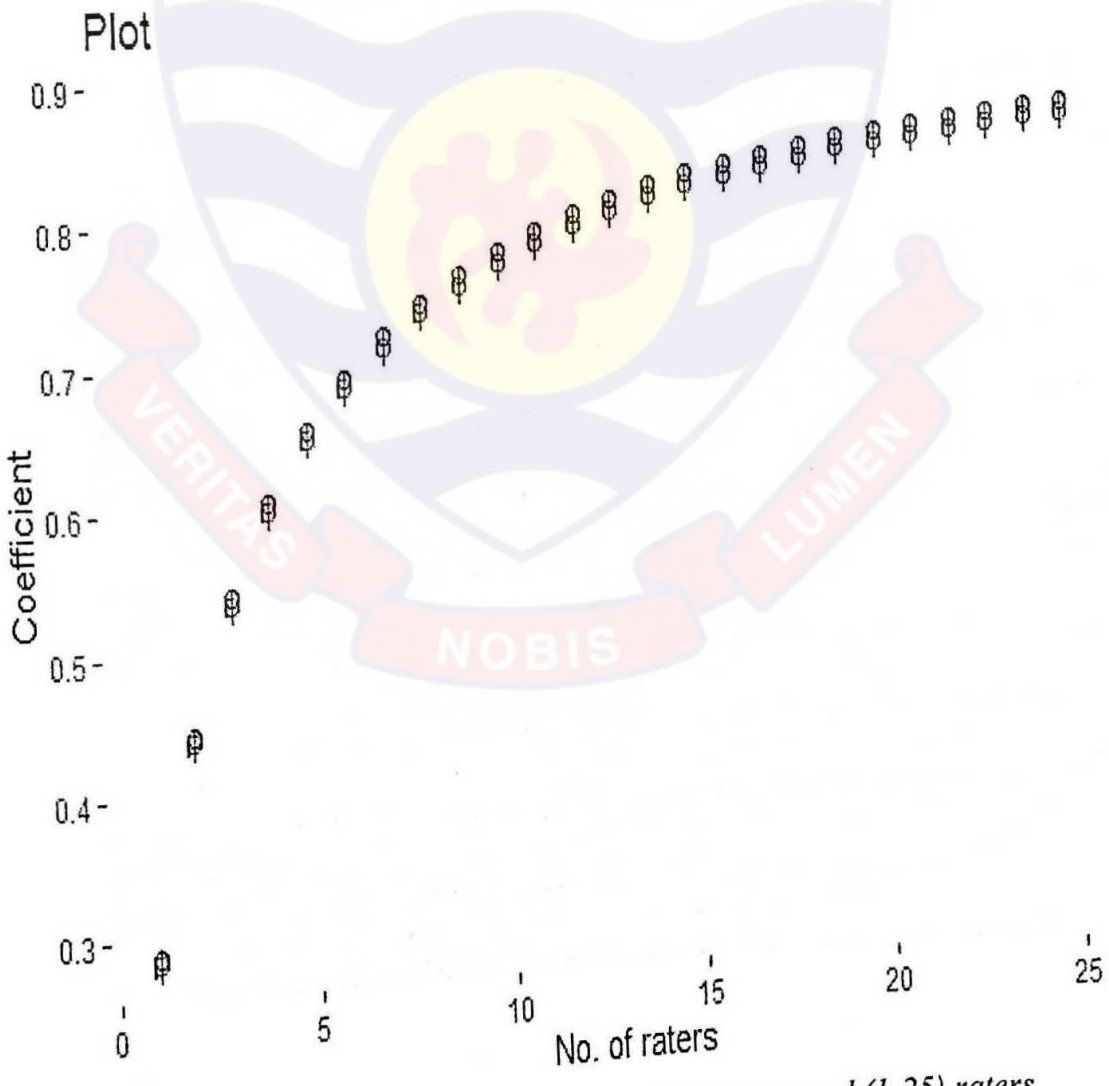


Figure 6- *Optimisation (Teaching) Graph with 12-items and (1-25) raters*

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

As shown in Figure 6, it can be observed that a very low-reliability index is produced (with 12-items) when there was a single rater (g<.30). Figure 6 showed that there is moderately established validity evidence when raters were ranging from 5 to 10. The results also showed sufficient validity evidence when there were small classes with 13 to 25 raters (with 12-items).

Research Question 4

How does non-response influence the validity of students' evaluation of courses and teaching?

This research question sought to understand how non-response affects the extent of dependability of students' evaluation of courses and teaching. First, the data were explored to understand the extent of non-response among the students. Afterwards, optimisation was conducted to examine how non-response can influence the quality of data provided by students concerning the evaluation of teaching and learning. With regards to the latter, the generalizability index and index of dependability together with their respective absolute and relative errors were examined to understand how non-response could affect the quality of the measurement.

The extent of non-response in student evaluation data

The data were analysed to explore the non-response patterns by comparing the expected responses and actual responses provided taking into consideration the levels of study. A ratio was computed for the expected and actual responses, and these ratios were compared across the levels. The ratios showed the extent of non-response such that a high ratio (closer to 1) showed a low level of non-response rate whereas a low ratio (far below from 1) depicted

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

a high level of non-response rates from the students. Tables 11 and 12, and Figure 7 present the details of the data.

Table 11- *Descriptive Statistics for Expected and Received Responses*

Level of Study	Expected responses	Received responses	Ratio
100	45722	27588	.60
200	48316	11945	.25
300	32692	13745	.42
400	21443	18113	.84
500	280	101	.36
600	149	141	.95
700	5	4	.80
800	3665	2121	.58
900	386	148	.38
Overall	152658	73906	.484

Source: Field Secondary Data (2021)

As shown in Table 11, a total of 152658 responses were expected from the entire student population. Out of this number, 73906 responses were received by DAPQA-UCC, which constituted a response rate of 48.4%. This suggests that more than half of the responses were not received (51.6%) indicating a high rate of non-response. Concerning specific levels of study, it was observed that non-response was more prevalent among level 200 (25%) students whereas non-response was less prevalent among level 600 students. This is presented in graphical form in Figure 7.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

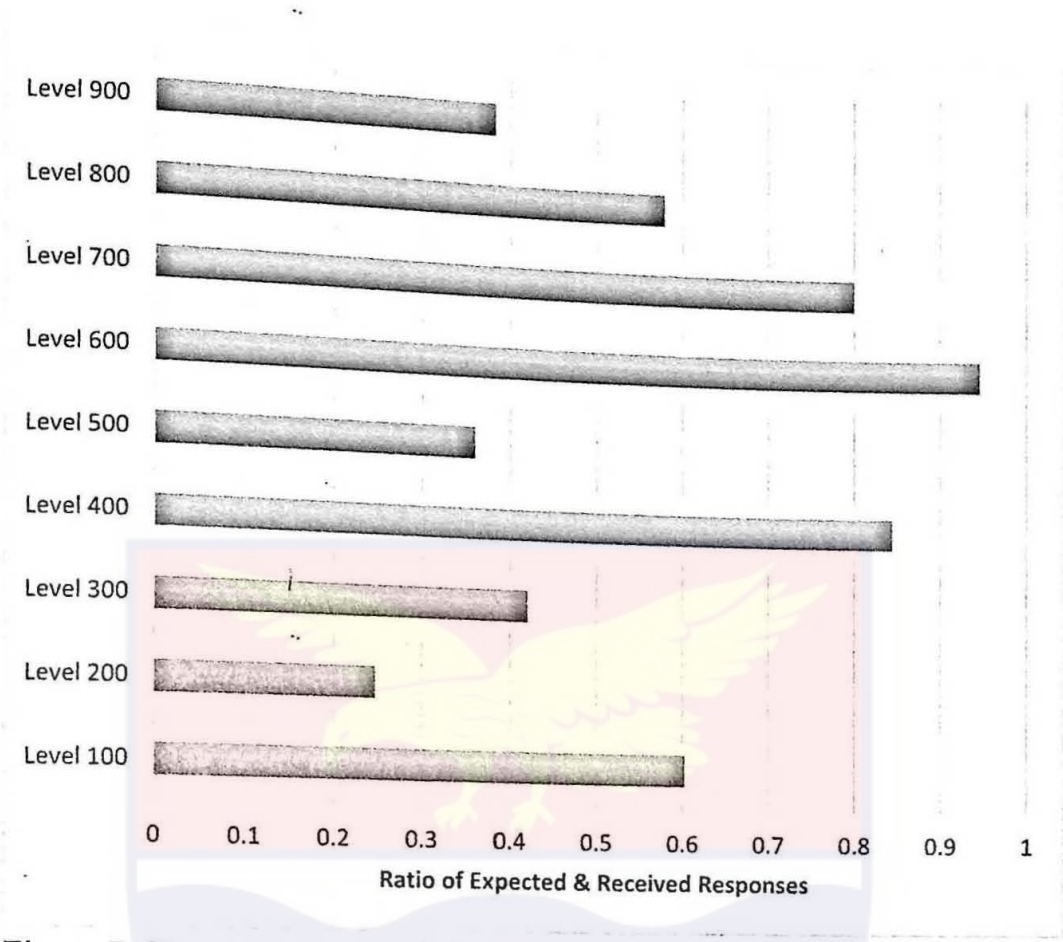


Figure 7- *Non-response graph for levels of study*
Source: Field Secondary Data (2021)

As shown in Figure 7, it can be observed that the non-response rates for levels 900, 500, 300 and 200 were below .50. Among the postgraduate students, the level 700 students had a high response rate, followed by the level 800 students. Non-response appeared to be more prevalent among the level 900 students (doctoral students).

Specific courses were selected and their response rates were explored. This was done to give much clarity to the results in Table 11. That is, the response ratio for some selected courses were computed and this has been presented in Table 12.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Table 12- *Response Ratio of Some Sampled Courses*

Course (Pseudonym)	Expected responses	Received responses	Ratio	% Response
ABO	3246	6	.001	.01
AFP	3265	7	.002	.02
ADE	3239	7	.002	.02
AGR	3315	11	.003	.03
AAW	97	1	.01	1.0
AKR	331	11	.03	3.0
ARQ	575	110	.19	19
ALT	358	91	.25	25
ACM	351	95	.27	27
AXI	665	176	.26	26

Source: Field Secondary Data (2021)

As displayed in Table 12, some of the courses had a response rate lower than 1% of the expected responses. For instance, a course like ABO, which had an expected response of 3246, recorded only 6 responses resulting in a .01% response rate. Courses AFP and ADE recorded response rates of .02% each. Whereas courses like AGR recorded a response rate of .03%, others also recorded response rates between 25 and 27 (e.g. ALT, ACM, AXI).

How non-response can affect the dependability of students' ratings

After determining the extent of non-response, optimisations were conducted to model how non-response can affect the data provided by students. Separate modelling was done for course appraisal data and teaching evaluation. The modelling was done by using non-response intervals of 1 (i.e., only one rater failed to respond), 10 (i.e, 10 raters failed to respond), and 47 (i.e., 47 raters failed to respond).

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Modelling non-response and course appraisal

Table 13, and Figures 8 and 9 present the results on how non-response can affect the extent of data provided by students regarding course appraisal. It must be noted that UCC uses 3-items as a proxy to the measurement of the quality of course contents.

Table 13- *Contribution of Non-response to Course Appraisal* (n_i=3)

No. of raters	Absolute error	Relative error	Phi	G
NR=10				
100	.025	.025	.643	.645
90	.026	.026	.634	.636
80	.026	.026	.623	.625
70	.028	.028	.610	.612
60	.031	.031	.593	.595
50	.031	.031	.571	.573
40	.035	.035	.541	.543
30	.041	.041	.497	.498
20	.055	.055	.428	.429
10	.095	.095	.302	.302
NR=47*				
245	.018	.018	.695	.697
198	.019	.019	.686	.688
151	.020	.020	.671	.674
104	.022	.022	.646	.648
57	.029	.029	.587	.589
10	.095	.095	.302	.302

NR- non-response; *Average non-response of all cases
Source: Field Secondary Data (2021)

The results, as shown in Table 13, revealed that a class of 100 students who engage in the appraisal of a course produced a generalizability index of .645 and a relative error of .025. It can be observed that the generalizability

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

index reduces to .636 if 10 of the students fail to respond to the evaluation. In instances where 20 students failed to participate in the survey, the generalizability index reduced from .645 to .625. In another case where 50 students failed to participate in the evaluation exercise, the reliability coefficient reduced from .645 to .573 and their corresponding errors also increased from .025 to .031. A similar trend is also observed when 10 out of 100 students participate in the evaluation exercise; the reliability index became .302 with a relative error of .095. Similar findings were also revealed when a non-response of 47 was used.

Figure 8 provides a trend analysis of the relative error against the frequency of non-response.

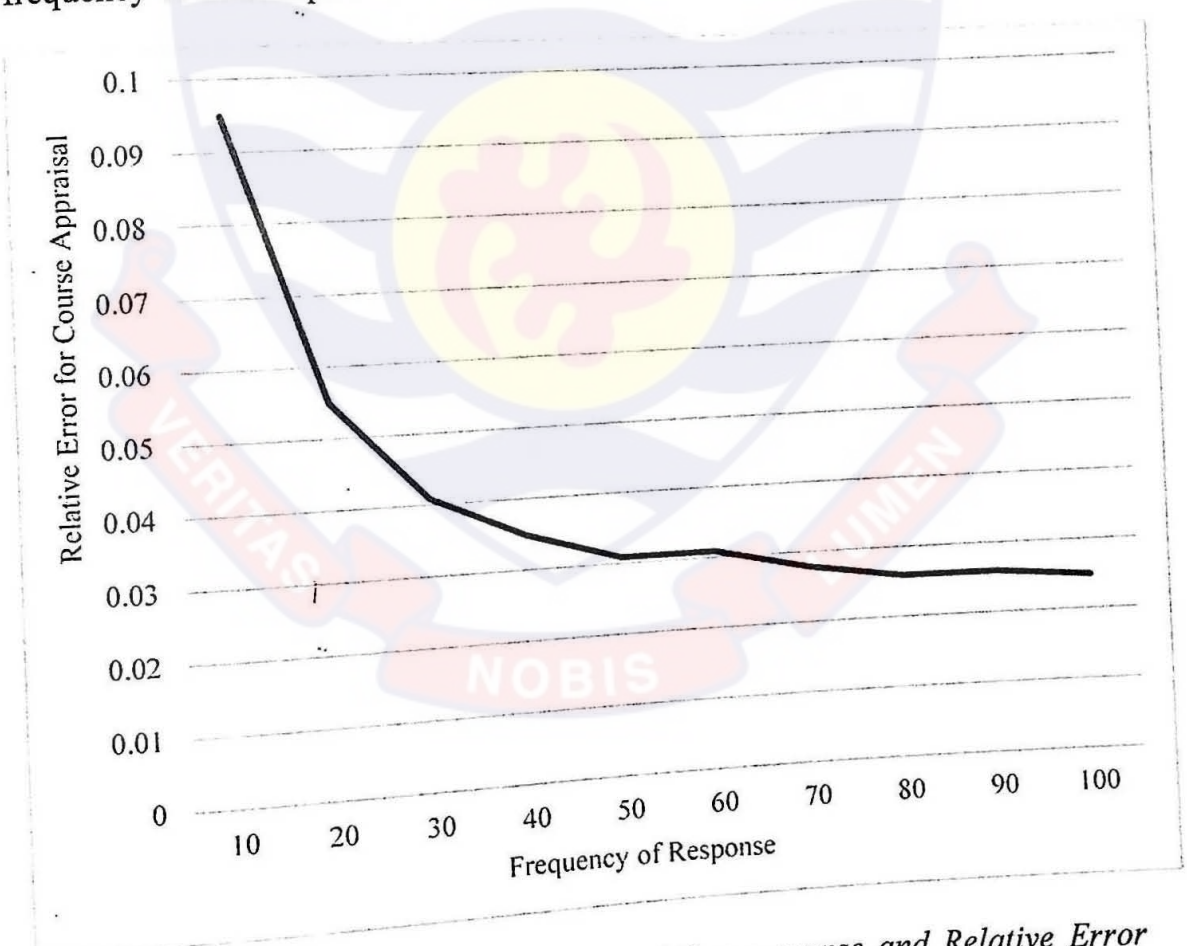


Figure 8- *Trend Analysis of Frequency of Non-response and Relative Error Variance for Course Appraisal*

Source: Field Secondary Data (2021)

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

It can be observed from Figure 8 that there appears to be a negative relationship between the frequency of responses and errors associated with the evaluation of courses. That is to say, that, increasing response rate results in decreasing relative error. As and when the number of responses decreases for course appraisal, the error relatively increases resulting in low dependability of the course appraisal data.

Figure 9 displays a trend analysis of how non-response is related to the generalizability coefficient and index of dependability.

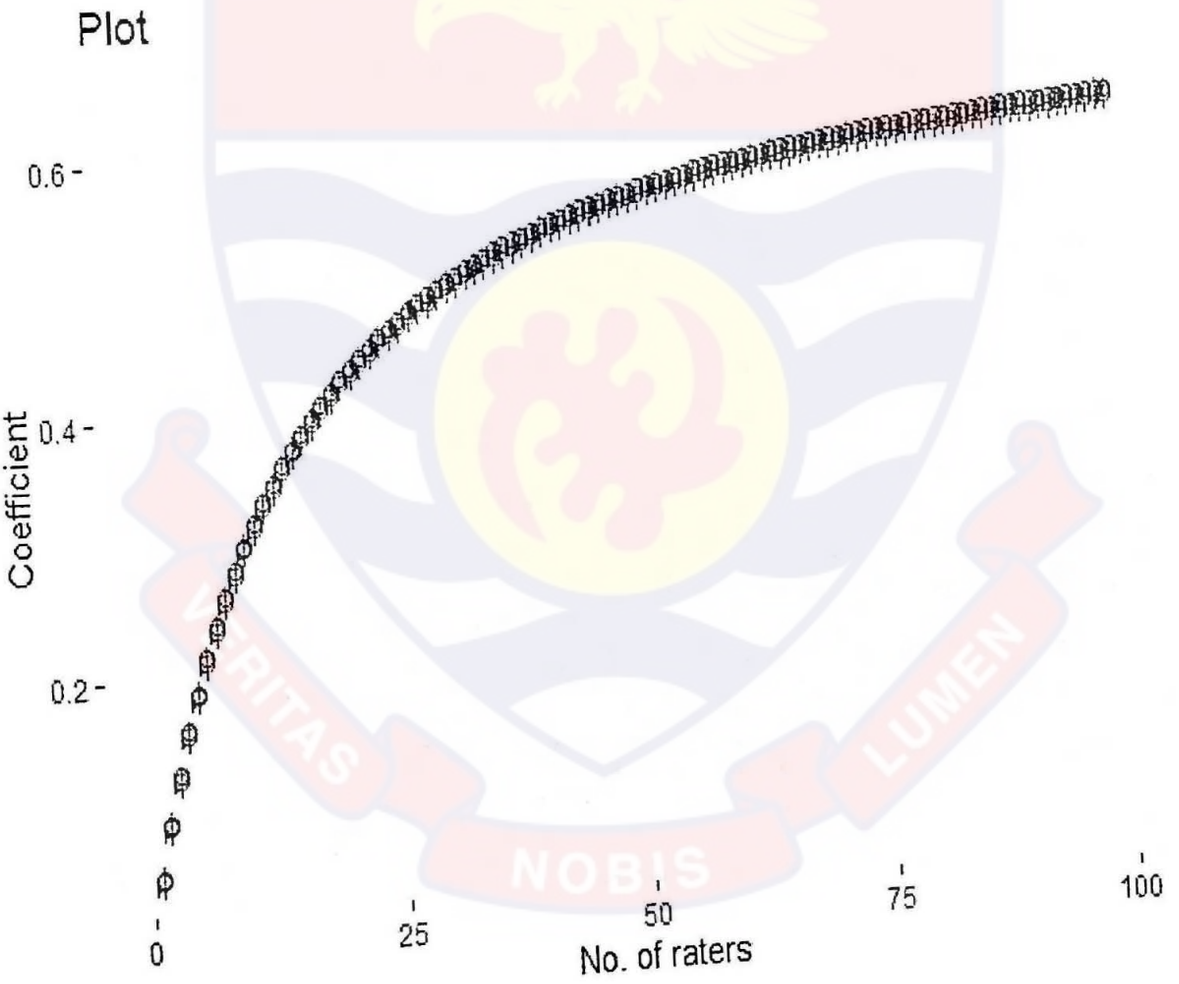


Figure 9- *Trend Analysis of Number of Raters and G/Phi Coefficient for Course Appraisal*

Source: Field Secondary Data (2021)

The trend analysis showed a positive relationship between the number of raters and the reliability coefficients (*g* and *phi*) regarding course appraisal.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

It can be observed that as the number of raters decreases from 100 to 0 along the x-axis, both reliability indices also reduces from .80 to 0 along the y-axis. In effect, non-responses reduce the number of expected raters and this further decreases the reliability coefficients.

Modelling non-response and teaching evaluation

Table 14, together with Figures 10 and 11 highlights how non-response affects the teaching evaluation data provided by students. It must be emphasised that 10-items were utilised for the evaluation of teaching exercise.

Table 14- *Influence of Non-response on Teaching Evaluation Data (n_i=10)*

No. of raters	Absolute error	Relative error	Phi	G
NR=10				
100	.011	.010	.932	.938
90	.011	.010	.929	.936
80	.012	.011	.926	.933
70	.013	.012	.923	.929
60	.013	.012	.918	.924
50	.015	.014	.911	.917
40	.017	.015	.901	.907
30	.020	.018	.885	.891
20	.026	.025	.854	.860
10	.044	.043	.774	.778
NR=47*				
236	.009	.008	.951	.951
189	.009	.008	.948	.948
142	.010	.009	.938	.944
95	.011	.010	.931	.937
48	.015	.014	.909	.915
1	.374	.373	.287	.288

*Average non-response of all cases
Source: Field Secondary Data (2021)

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

The results as presented in Table 14 revealed that a class of 100 students in UCC will produce a generalizability index of .938 and an associated relative error of .010. If, for example, 30 students in the class failed to participate in the evaluation exercise, the reliability coefficient will reduce to .929 and the relative error will increase to.012. A similar trend of results was revealed such that the more students failed to respond to the evaluation survey, the more the reliability coefficients reduced and errors increased. This is also reflected in Figures 10 and 11.

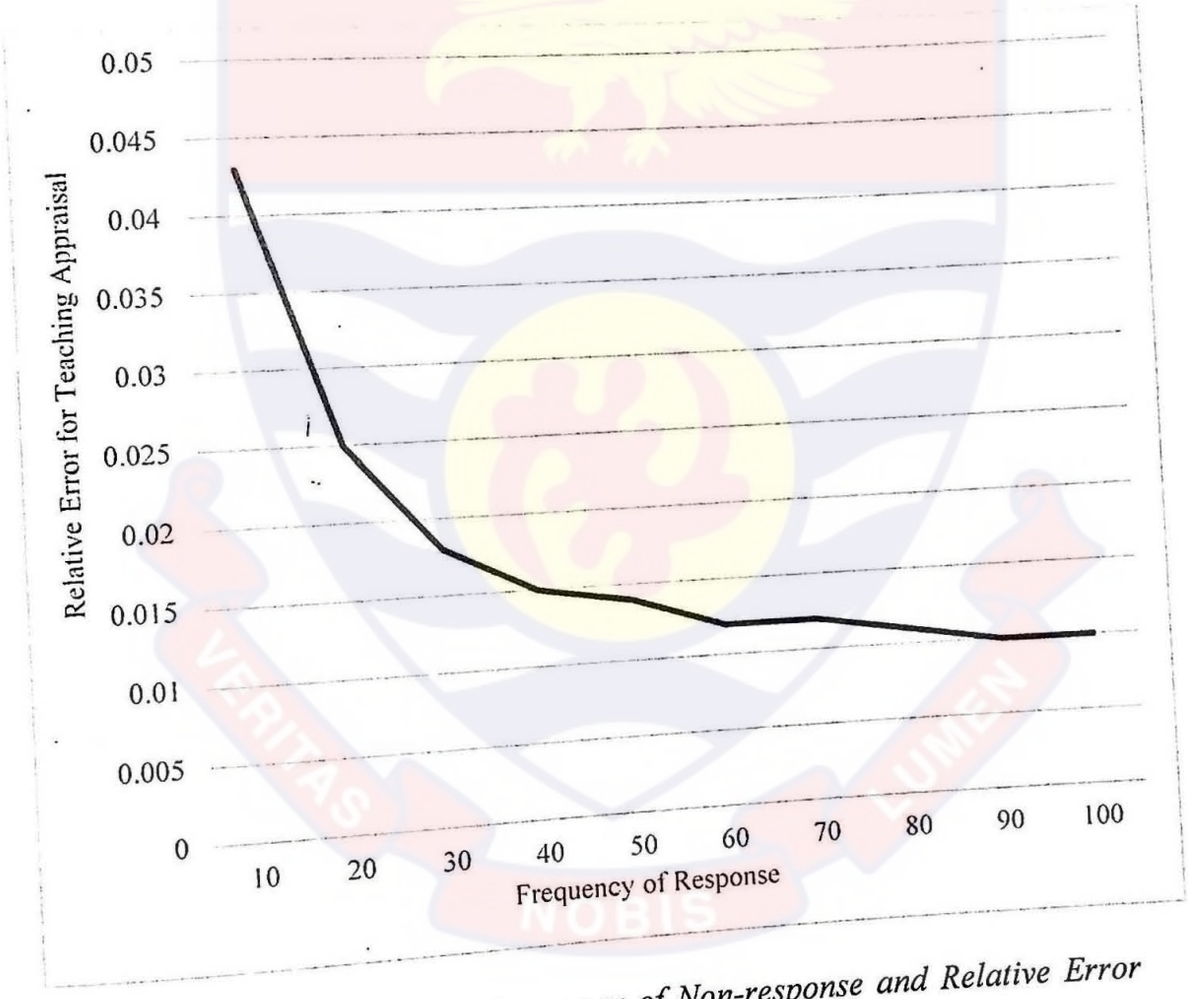


Figure 10-*Trend Analysis of Frequency of Non-response and Relative Error Variance for Appraisal of Teaching*
Source: Field Secondary Data (2021)

From Figure 10, it can be observed that there is a negative relationship between the frequency of responses and errors associated with rating variability in teaching evaluation. That is, increasing responses results in decreased relative

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

error. As and when the number of responses decreases for teaching evaluation appraisal, the error relatively increases resulting in low dependability of the teaching evaluation data provided by students.

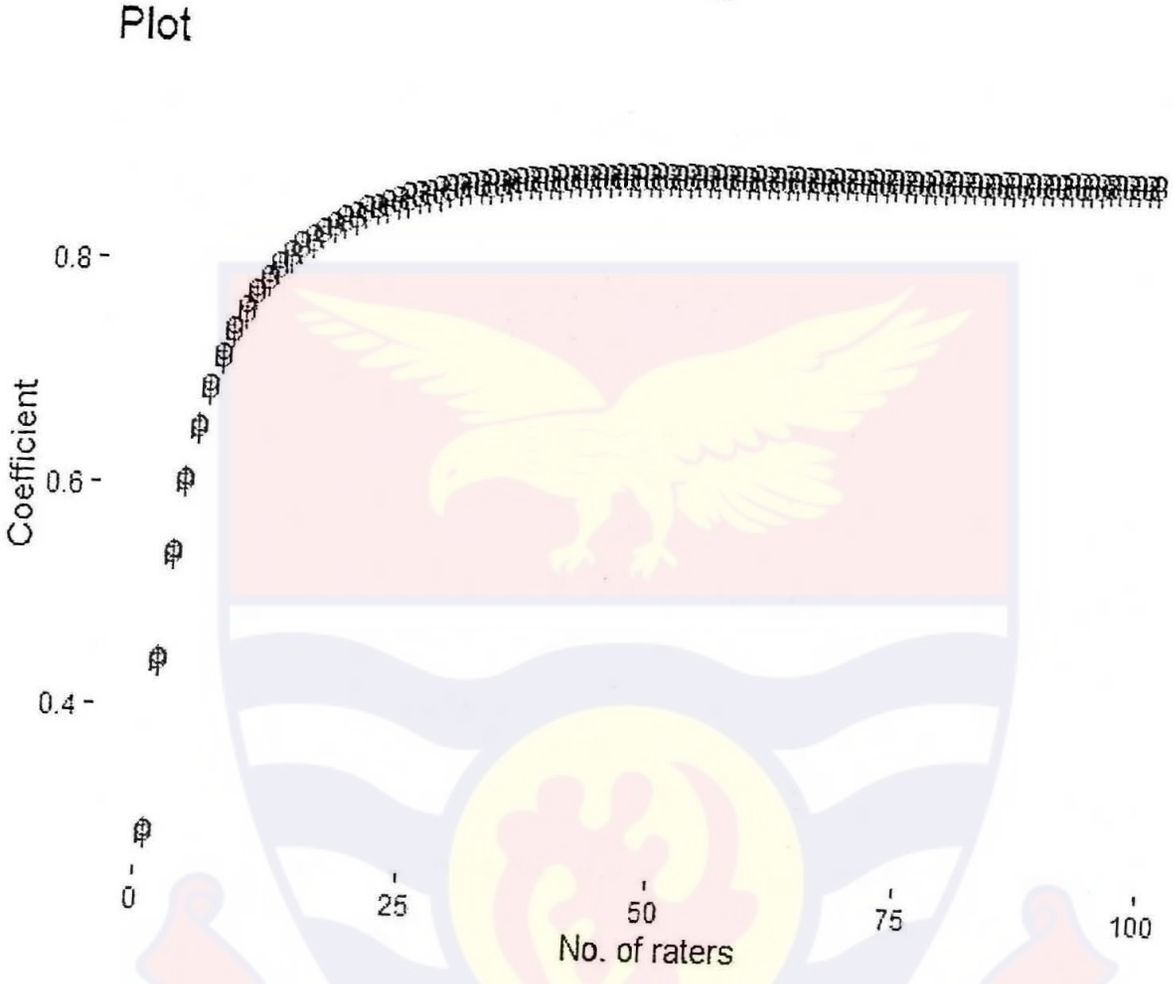


Figure 11- *Trend Analysis of Number of Raters and G/Phi Coefficient for Appraisal of Teaching*

Source: Field Secondary Data (2021)

As displayed in Figure 11, there appear to be a positive association between the number of raters and the reliability coefficients (both *g* and *phi*) regarding teaching evaluation. It can be observed that as the number of raters decreases from 100 to 0 along the x-axis, both reliability indices also reduces from .80 to 0 along the y-axis. In effect, non-responses reduce the number of expected raters and this further decreases the reliability coefficients.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

error. As and when the number of responses decreases for teaching evaluation appraisal, the error relatively increases resulting in low dependability of the teaching evaluation data provided by students.

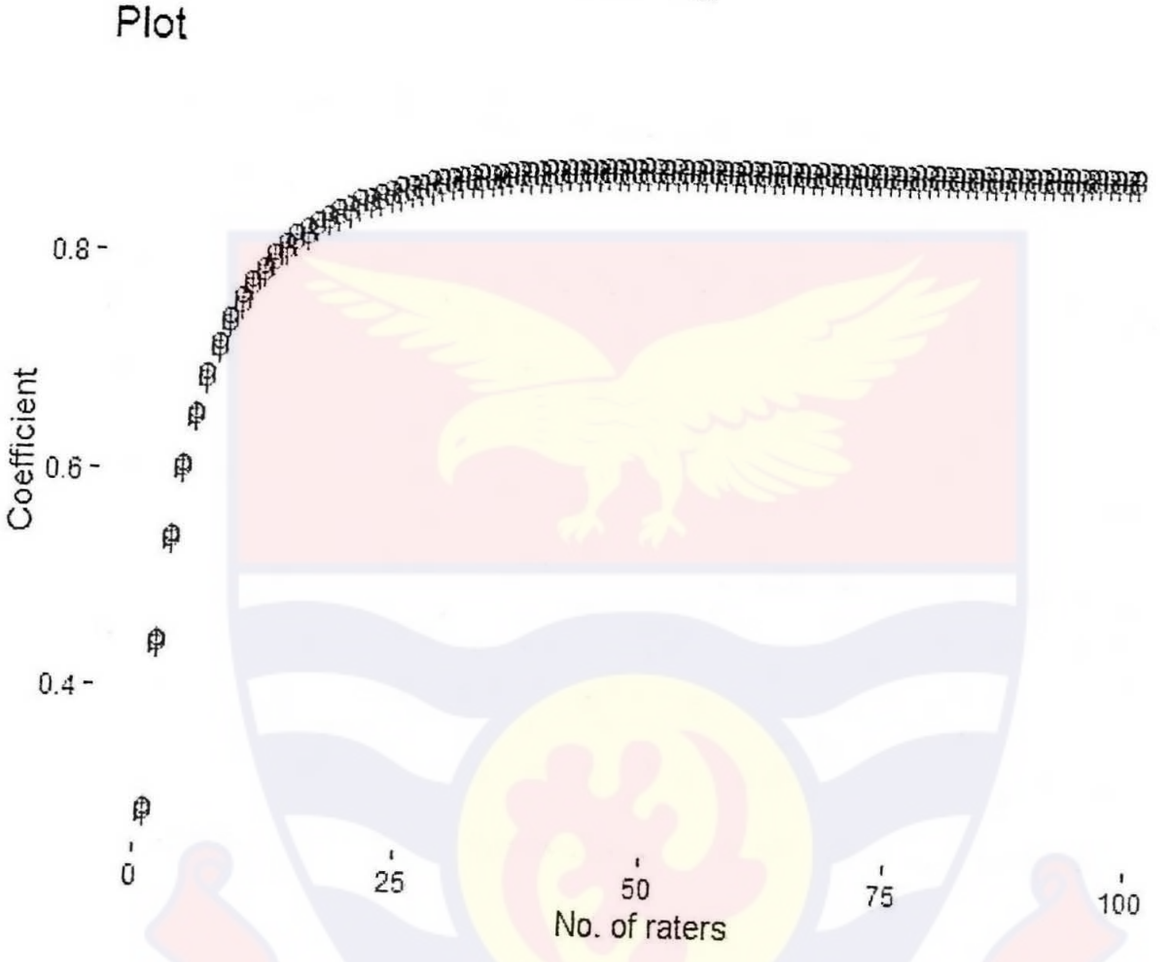


Figure 11- *Trend Analysis of Number of Raters and G/Phi Coefficient for Appraisal of Teaching*
Source: Field Secondary Data (2021)

As displayed in Figure 11, there appear to be a positive association between the number of raters and the reliability coefficients (both *g* and *phi*) regarding teaching evaluation. It can be observed that as the number of raters decreases from 100 to 0 along the x-axis, both reliability indices also reduces from .80 to 0 along the y-axis. In effect, non-responses reduce the number of expected raters and this further decreases the reliability coefficients.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Research Question 5

What are the behaviours of students (halo effect and severity effect) in the appraisal of courses and teaching?

This research question evaluated the rating behaviours of students during the appraisal of courses and teaching. Details of the results are shown in Tables 15 to 17.

Table 15- *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
Source: Field Secondary Data (2021)

Table 15 provides the summary of fit statistics for the rating behaviours of students in evaluating teaching and course contents. Although both infit and outfit statistics are provided, more emphasis will be placed on infit statistics since the outfit statistics are sensitive to outliers. The results (in Table 15) revealed that about 34.8% of the students had an acceptable fit. This 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 a halo effect. The misfit statistics showed that there was more variance in the ratings than expected

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

among the 34.7% of the students. This indicated that these students were inconsistent and unpredictable

Table 16 further provides a summary of the rater measurement report which presents details of the entire model.

Table 16- *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
Source: Field Secondary Data (2021)

The summary statistics included a model summary and location summary of the analysis (Table 16). A closer look at the model summary showed a significant chi-square value, χ^2 (213) =258.1, p<.001. The chi-square tests the hypothesis that all student raters were equivalent in their level of severity. Based on the chi-square test, the results showed that 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 teaching and course contents. This was also confirmed by the moderate reliability of the separation coefficient (which was further from zero) and a separation ratio (which was slightly greater than 1).

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

The location summary, as shown in Table 16, 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 rating and high variances in the ratings.

In Table 17, individual cases with their detailed parameters were studied. This comprised those raters who recorded misfit as well as those who recorded overfit.

Table 17- *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
Source: Field Secondary Data (2021)

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

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 17, revealed that those raters who were misfitting consistently provided low ratings than expected. Rater 12, for instance, gave the instructor a score of 2.62 instead of a true/expected score of 2.77. Similarly, rater 231 rated an instructor with a score of 2.31 instead of an expected score of 2.60. Raters 15, 6, and 94 were all found following a similar trend; providing a score lower than the expected score of the instructor. One thing was also common among the raters who were misfitting, that is, they were all found at the higher 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 at 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 score 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.

Research Question 6

How effective are evaluation items in the measurement of the quality of course content and teaching?

This research question examined how effective the evaluation items were in the measurement of the quality of course content and teaching. Three different aspects of the items were assessed whether: (1) the structure of the item was appropriate and clear enough, (2) the criteria was well understood and the items accurately measured a particular criterion, (3) the rating scale is of

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

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 18 presents the details of the result of the unexpected responses.

Table 18- *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
Source: Field Secondary Data (2021)

The results presented in Table 18 revealed that items 8, 6, 5, 1, 9, and 3 had problems during the evaluation exercise. These items were further investigated to identify the problem at hand. Also, the criteria mode of delivery (teaching quality) was flagged as problematic and thus, there is the need for further investigation (Table 18). The rating scale (especially the first/starting point of the scale) also had challenges and posed threat to the accuracy of data obtained.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Specific Item Measurement Report

Table 19 presents the summary statistics for the item facet.

Table 19- *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
Source: Field Secondary Data (2021)

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

The model summary in Table 19 showed a significant chi-square test, $\chi^2(12) = 72.1$, $p < .001$. The chi-square tests the hypothesis that the items are of equal difficulty. The outcome of the test suggests that the items were not of equal difficulty. The separation strata index of 3.09 implies that the items have 3 statistical distinct levels of difficulty. This signal multi-dimensionality of the instrument/scale used with three dimensions, which is more than the two dimensions used in the study (i.e., mode of delivery and course contents). The separation ratio of 2.07 and reliability of .81 (values close to 0 is preferred) provided much evidence to the fact that the heterogeneity among the items was more than expected.

As presented in Table 19, 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 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 misfit (items 3, 4, 8, and 13). The items which recorded overfit meant that they were redundant, such that these items failed to provide new information which adds 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 a single measurement construct.

It must be mentioned that items that recorded poor fit (whether overfit or misfit) appear problematic and, therefore, needs to be either revised or discarded. These items which recorded poor fit were selected and studied in this study. These items were evaluated to find out what could be wrong with them

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

or what is making such items less effective as indicated by the results in Table 19.

Evaluation of items flagged as unclear or failed to measure the construct

Item 3: The lecturer effectively communicated what he/she was trying to teach

Response format: Not very well -> Not well -> Well -> Very well

This item was factored under the "mode of delivery" section of the evaluation form indicating that it measured specific aspect of the teaching behaviours of the lecturers. A critical look at the item showed that the item is relatively broad in terms of what it is supposed to measure. The use of the word "*effectively*" seems to suggest that there are specific standards against which the actual behaviour is being measured. When something is effective, it means that that thing has an expected outcome. In this case, the item is looking at whether the lecturer effectively communicated what he/she was trying to teach. From the view of the researcher, this statement is a good proxy for the measurement of the construct/trait (i.e., teaching ability).

The item, however, appears unclear and raises several concerns: What standards are the students evaluating the lecturer's delivery against? What behaviours of the lecturer signals that this particular instructor communicated effectively what he/she was teaching and to what extent? At what point should a lecturer be rated as '*not very well*'? At what point should a lecturer be given a rating as '*not well*'? Until the rater (i.e., student) is clear about these questions, such an item may be unclear. Such an item needs to be revised or expanded to have several aspects to improve clarity.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Item 4: The lecturer used class time to fully promote teaching and learning

Response format: Not very well -> Not well -> Well -> Very well

Item 4 attempted to find out whether lecturers used class time to fully promote teaching and learning and as such, the item was placed under the 'mode of delivery section of the evaluation instrument. This item appears too broad and may constitute several pieces of behaviours to be demonstrated. The researcher believes that several pieces of behaviours characterise instructors who use their class time to fully promote teaching and learning. How many of them should be present for a rater to rate '*well*' or '*very well*'? Is one of these behaviours adequate to earn a rating label '*well*'? How often should this behaviour occur to earn a rating label of '*well*'? When should a lecturer be rated '*not very well*' in terms of promoting teaching and learning in class?

Also, the use of the word '*fully*' in the statement may contribute to errors in the evaluation exercise. The word 'fully' may suggest entirely or completely or to the fullest extent. This connotes that there is a measure or benchmark which should be entirely achieved. What are these benchmarks against which the raters are required to provide their rating? At what point can it be said that a lecturer has *fully* promoted teaching and learning? At what point can it be also mentioned that an instructor has *not fully* promoted teaching and learning? Issues of this nature may breed subjectivity on the part of the raters where each rater interprets the item differently. Such an item must be revised or expanded to have several aspects to improve clarity.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Item 6: The lecturer encouraged students to analyse issues from different perspectives

Response format: Not very well -> Not well -> Well -> Very well

Item 6 sought to examine whether the lecturers encouraged students to analyse issues from different perspectives. This item was placed under the sub-section 'mode of delivery'. From the researcher's expertise, this item qualifies to form part of the items measuring the teaching behaviours of lecturers. This item, however, may be unclear to the raters, especially when using this response format (i.e., *Not very well -> Not well -> Well -> Very well*). What does it mean to analyse issues from different perspectives? Does this include a lecturer who is accommodative to all responses from different students? Or it is about thinking outside the box? What should be looked at before a lecturer is rated with a label of '*very well*' or even '*not well*'. This item may be subjective and obviously will elicit diverse meanings from the raters.

Item 8: The lecturer made time during lectures for questions and discussions

Response format: Not very well -> Not well -> Well -> Very well

Item 8 solicited information on whether the lecturer made time during lectures for questions and discussions. This item was also placed under the sub-section of the evaluation form labelled 'mode of teaching'. This item is placed appropriately, at least, looking at the face value and as such, can provide a meaningful measure of lecturers' mode of teaching. Nevertheless, this item appears to be double-barrelled. This means that the item has another question embedded in it. Looking at the item, there are two issues: (1) the lecturer made time during lectures for questions; (2) the lecturer made time during lectures for discussions. Raters may be confused about which statement to respond to. How

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

should students' rate when a lecturer make time for questions but this does not necessarily end up in discussion? How should such a lecturer be rated? Issues like this result in subjectivity and encourage heterogenous rating behaviours. This item needs to be revised and split into two different items.

Item 13: The course content was detailed (i.e., each week with its specific content, materials to read, etc)

Response format: Not detailed -> Slightly detailed -> Detailed -> Very detailed

Item 13 required students to indicate whether the course content was detailed (i.e., each week with its specific content, materials to read, etc). This item was placed under the sub-section of the evaluation form labelled 'course content'. The statement reflects the construct being measured. What appears wrong with the item is that it is loaded with several indicators. How should the rating be done if the course content is detailed where each week comes with its specific content but the materials to read is not detailed? Should the rater provide a rating label as '*slightly detailed*' or '*detailed*'? Also, the use of 'etc' in the statement may not be appropriate since this leaves the rater hanging with regards to what is meant by that. Such an item should be revised or expanded to accommodate the several aspects intended to measure to improve clarity.

Evaluation of items flagged as redundant when measuring the construct

Item 1: The lecturer demonstrated knowledge of the subject matter

Response format: Not very well -> Not well -> Well -> Very well

This item elicited information regarding whether the lecturer demonstrated knowledge of the subject matter. This item was placed under the 'mode of teaching' sub-section of the evaluation form which appears appropriate. The item was flagged as redundant indicating that it did not

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

measure anything new apart from what other items have already measured. From the viewpoint of the researcher, knowledge of the subject matter can appropriately and accurately be determined by persons who have mastery in the subject area. In this context, students may not be in the best position to rate the subject matter of the lecturer since this will require the students to have much knowledge in the course the lecturer is handling. Additionally, students will be required to be able to know when to rate a lecturer's subject matter knowledge as 'not very well' or 'not very'? Indeed, this may require some expertise and this may render the item 'impotent' in terms of measuring the construct. As a result, students may rate inconsistently on this item and thus, the item needs to be revised or discarded.

Item 5: The lecturer encouraged students to read materials beyond the suggested readings and lecture notes

Response format: Not very well -> Not well -> Well -> Very well

In item 15, the focus was to examine whether the lecturer encouraged students to read materials beyond the suggested readings and lecture notes. This item was placed under the sub-section 'mode of teaching' of the evaluation form. From the researcher's viewpoint, this item seems misplaced because just the lecturer encouraged students to read materials beyond the suggested readings and lecture notes do not directly reflect the quality of the mode of teaching and do not guarantee an excellent mode of teaching. The lecturer may encourage students to read beyond the suggested readings and lecture notes, but an unsuitable mode of teaching will be adopted. This item needs revision or replacement.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Item 10: The lecturer was concerned with the academic progress of students

Response format: Not very well -> Not well -> Well -> Very well

Item 10 sought to examine whether the lecturer was concerned with the academic progress of students, and the item was placed under the 'mode of teaching' section of the evaluation form. This item was flagged redundant possibly because it does not appear to measure the mode of teaching of the lecturer. A lecturer who is concerned with the academic progress of students does not necessarily translate into a quality/effective mode of teaching. Thus, a lecture may be concern about the academic progress of students and at the same adopt inappropriate/ineffective mode of teaching during engagement with students. This question is more of a construct irrelevance item and does not directly measure the construct under investigation. This item needs replacement or revision to reflect the construct.

Criteria Measurement Report

The criteria measurement report provides information on the role played by the sub-dimensions of the instrument. In this study, two sub-dimensions were the focus of the study: mode of teaching and course contents. The report in this section seeks to assess whether each of the scales functioned as it was supposed to. Issues like were the items under specific sub-dimension measuring the construct in question, or some sub-dimensions were redundant. This is a follow-up to the summary of the item report shown in Table 18 which found traces of criterion disfunction. Table 20 presents the details of the results on the criteria.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Table 20- *Criteria Measurement Report*

	Criteria/Sub-dimensions	
	Mode of Delivery (Teaching Appraisal)	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
Source: Field Secondary Data (2021)

The results on the mode of delivery criterion, as presented in Table 20, showed a significant chi-square test which tests the hypothesis that the items (under the mode of teaching) 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 (in Table 20) on the course content dimension revealed a non-significant chi-square test which tests the hypothesis that the items (under

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

	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

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

	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
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

Source: Field Secondary Data (2021)

The results in Table 21 showed that the majority of the raters were not able to appropriately use the scale. Taking rater 6, for instance, scale category 1 recorded very low frequency count and percentage (n=1, 8%) indicating that 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 had a logit measure of .60, category 3 had.39

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

and category 4 had a logit measure of .51. It can be observed that instead of the average logit measure increase 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 unclear to them. This is evident from the graphs. Taking rater 43, for example, the rater 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 category 1 and 2 (see Table 21).

Figures 12-15 presents the details of the probability curves for some selected students. Other samples of probability curves have been also shown in Appendix F.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

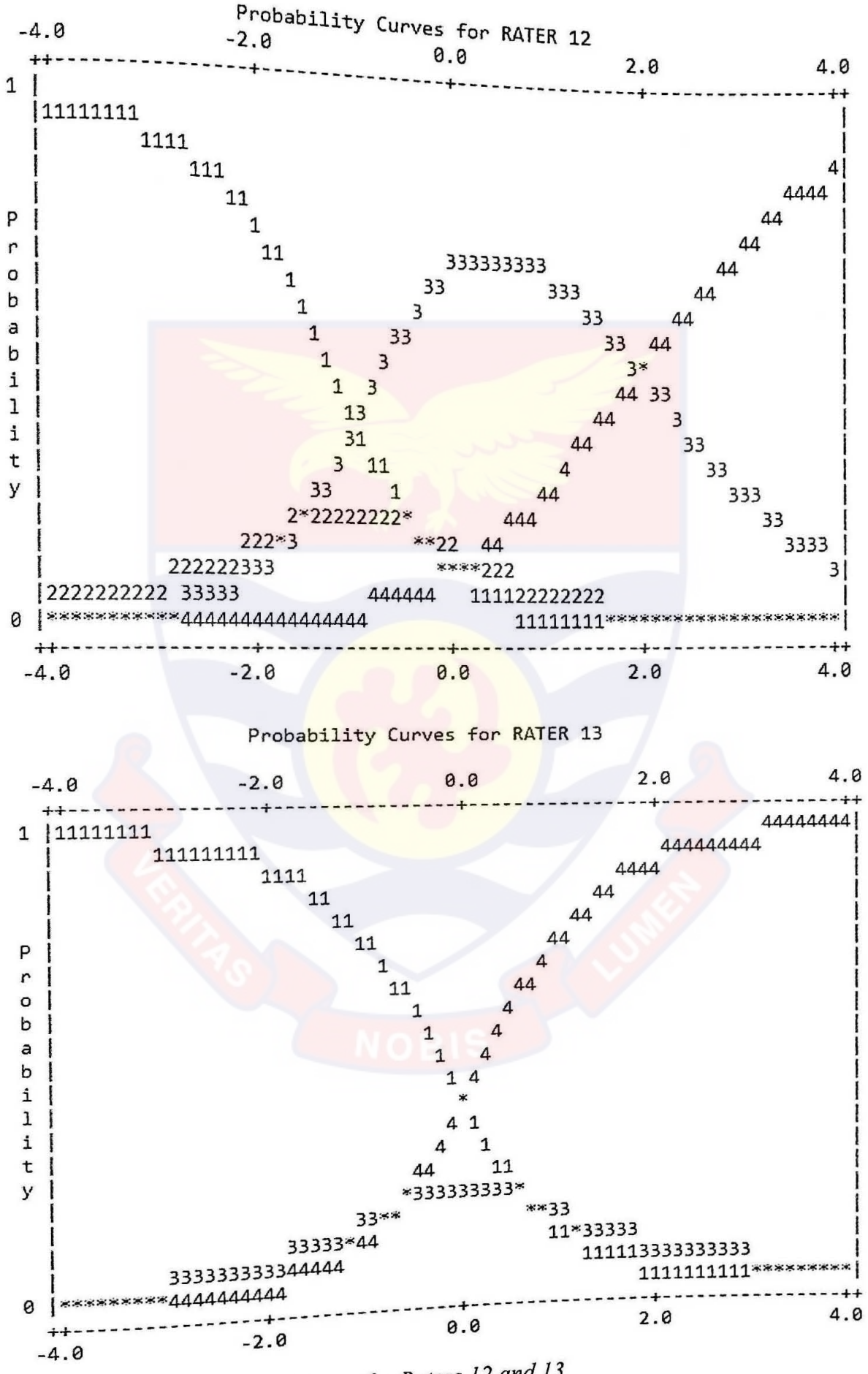


Figure 12- *Probability Curves for Raters 12 and 13*

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Figure 12 showed the probability curves for raters 12 and 13 based on their responses. A closer look at the probability curve for rater 12 revealed that this particular rater did not have an adequate understanding of how to use the rating category of 2; this category did not have a peak. This showed that rater 12 did not accurately assign a score using the scale category of 2 for those lecturers whose performance merited that scale category. As shown in Figure 12, rater 13 showed a poor understanding and use of scale category 2 and 3. Rater 13 over-used the extreme parts of the scale category and thus, did not discriminate between lecturers performance/ability which deserved a score of 2 and 3 (see Figure 12).

Figure 13 shows the probability curves for raters 43 and 60. Both raters found it difficult to use the scale category 1 indicating that they could not discriminate between the use of scale category 1 and 2. It appeared that these raters scored lecturers who were supposed to obtain a score of 1 were instead scored 2. A critical look at the probability curve for rater 60 indicates that this particular rater provided a score of 2 for the lecturers with the least ability. A similar trend of results was found in rater 43.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

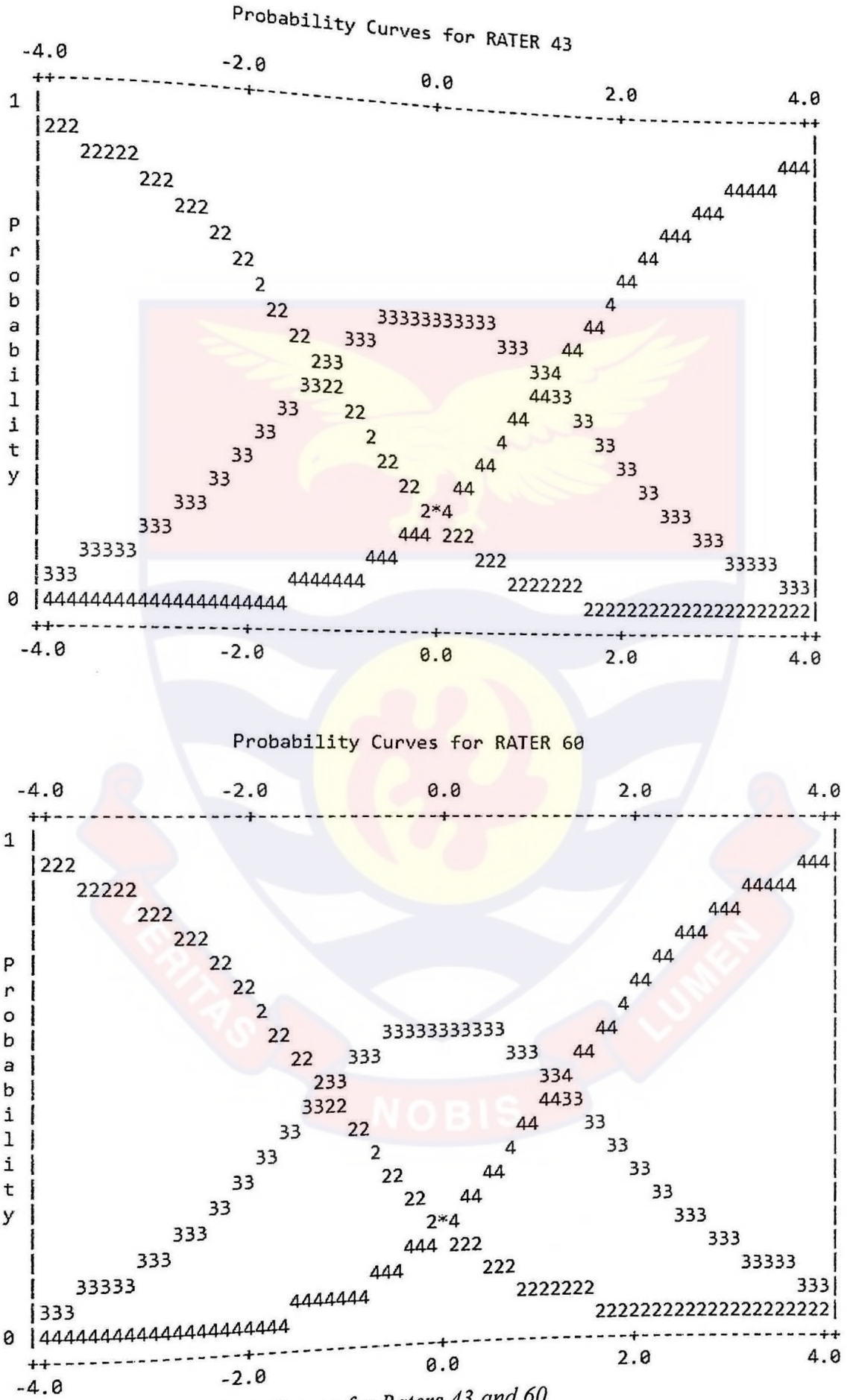


Figure 13- *Probability Curves for Raters 43 and 60*

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

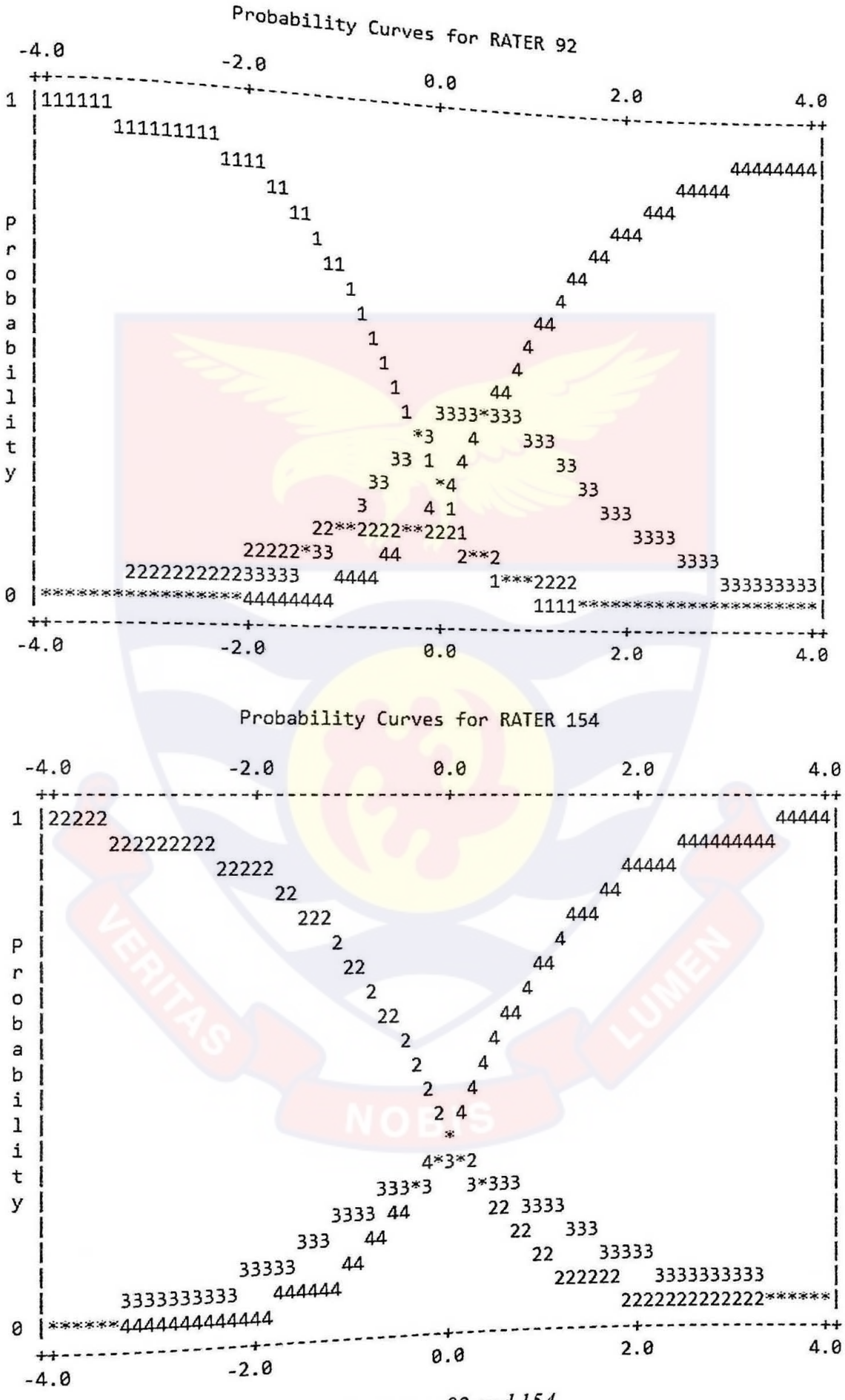


Figure 14- *Probability Curves for Raters 92 and 154*

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

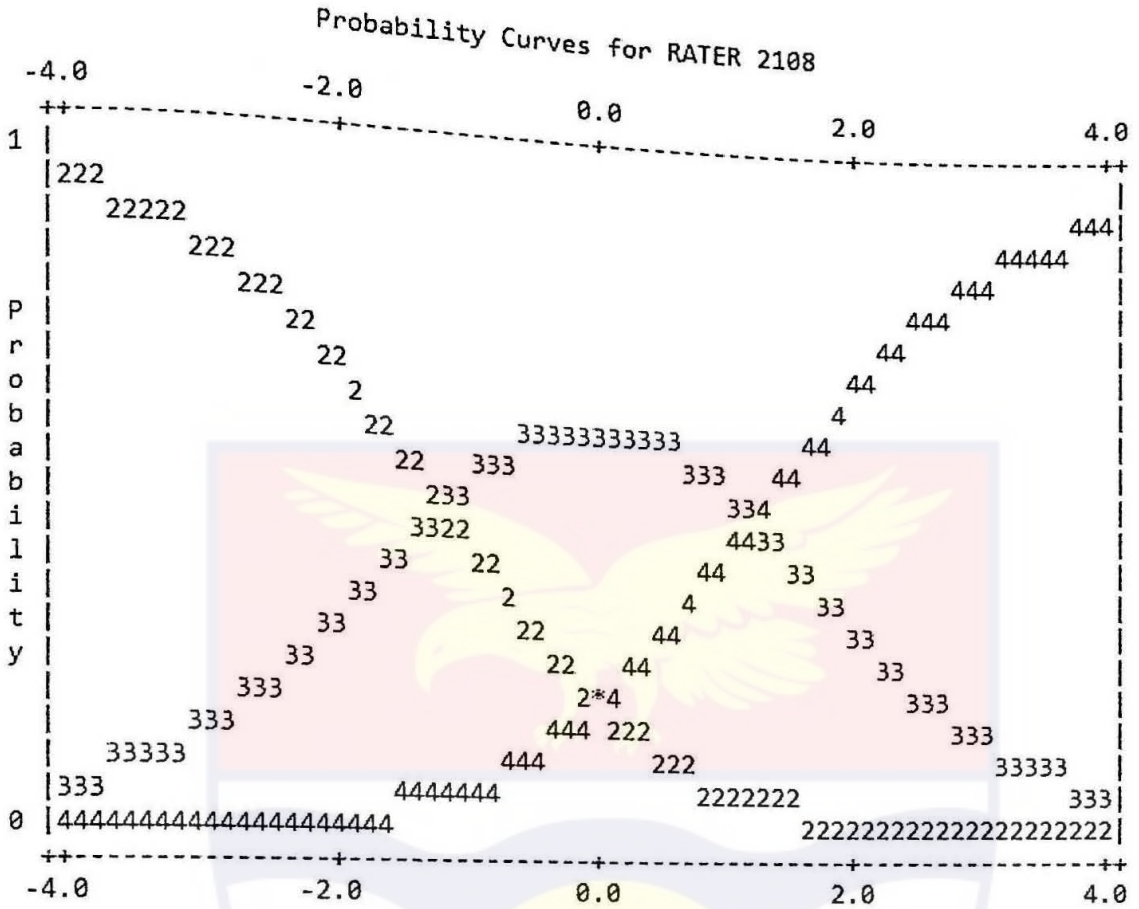
As displayed in Figure 14, the probability curve for rater 92 revealed a poor use of the rating scale. For this particular rater, the scale category 2 was not used and the scale category appeared to be misunderstood. That is, the rater was not able to discriminate between the scale categories 1 and 2, as lecturers with an ability that merited a score of 2 were given a score of 1. Also, rater 92 overly used the extreme scores on the scale category.

Rater 154 also showed poor use and understanding of the rating scale. The scale categories 1 and 3 were not used and this implied that rater 154 did not have an adequate understanding of the use of those scale categories (see Figure 14). It was observed that the rater was not able to discriminate between performances which required a score of 1 and 2, as well as 3 and 4. To explain further, all lecturers who functioned at the lower part of the ability scale were more likely to be given a score of 2 whereas all lecturers who performed at the higher part of the ability scale were more likely to be awarded a score of 4.

The scale category 1 was also found to be unclear for raters 2108 and 2613. Referring to Figure 15, it is clear that the probability of a lecturer obtaining a score of 1 was very low (negligible) such that those lecturers with the least performance were more likely to be awarded a score of 1 instead of 2. This was the same for both raters as shown in Figure 15.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui



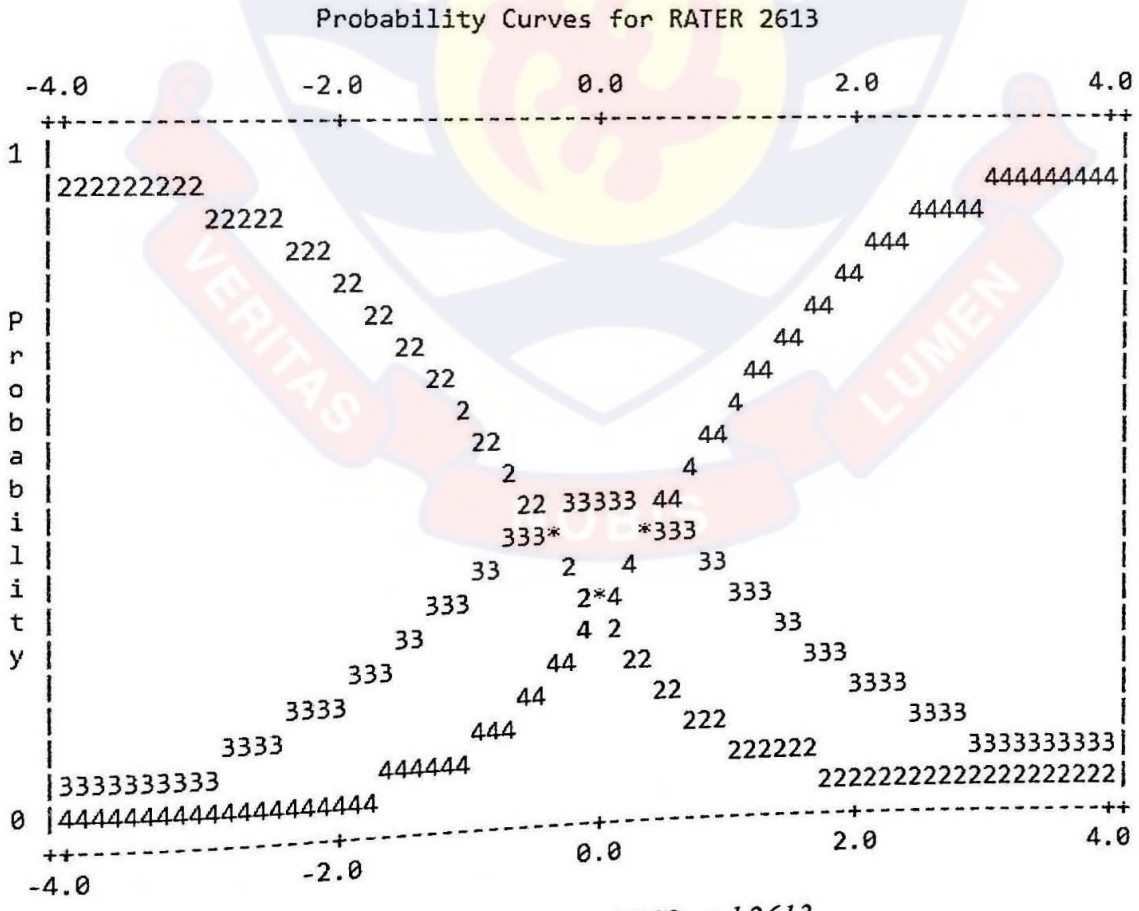


Figure 15- *Probability Curves for Raters 2108 and 2613*

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Research Question 7

How fair are lecturers rated in terms of the quality of lecturers' teaching?

This research question assessed the extent to which lecturers in UCC were fairly rated by the students regarding the appraisal of courses and teaching. The details of the results are shown in Tables 22 to 24.

Table 22- *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%)

Source: Field Secondary Data (2021)

The result in Table 22 highlights the fit statistics for instructors/lecturers which provides information on whether lecturers ability was 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 ability. The majority of the instructors were misfitting (47.6%) and overfitting (20%) suggesting that the ability of the lecturers was not measured properly or the description given to the course contents were far from what exists in reality.

Table 23 further provides a summary of the person (lecturer) measurement report which presents details of the entire model.

Table 23- *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

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

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
Source: Field Secondary Data (2021)

The results, shown in Table 23 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' ability was not accurately measured.

Some of the specific cases of poor fit (misfit) were selected and studied. This is presented in Table 24.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Table 24- *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
Source: Field Secondary Data (2021)

As shown in Table 24, 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 indicate 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.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Research Question 8

What are the biases in students' appraisal of courses and lecturers' teaching?

This research question evaluated the biases in students' appraisal of course and lecturers' teaching. Two different interaction analyses were conducted; (1) instructor-by-item interaction and (2) instructor-by-rater-by-item interaction. Table 25 presents the details of the analysis.

Table 25- *Bias Interaction for Course and Teaching Appraisal*

Statistic	Instructor-by-Item	Instructor-by-Rater-by-Item
Bias size (Mean) ± S.E	.00±.62	-.49± 1.81
Min. bias size± S.E	.14±1.64	-.01± 1.52
Max. bias size± S.E	-1.74± 1.64	-2.80± 1.94
Min. *t*-values	-.01	-.04
Mean *t*-values	-.03	-.28
Max. *t*-values	-2.25	-1.70
% $t \geq 2$	.004%	0
Min. *t*(df)	1	1
Max. *t*(df)	19	1
Chi-square	121	707
p-value	.70	.99

Source: Field Secondary Data (2021)

The results, as presented in Table 25, did not find any evidence of a bias interaction in the data. Taking the instructor-by-item interaction, for example, the bias size was very low (*bias*=.00) and thus, the percentage of *t*-values that were greater or equal to 2 was as low as .004%. The minimum *t*-value was -.01

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

and the maximum was -2.25. The chi-square test was non-significant indicating a non-significant instructor-by-item interaction.

The results on the instructor-by-rater-by-item interaction also revealed little evidence of interactions among the facets. The chi-square test showed a non-significant result suggesting that there is no significant interaction effect for instructor-by-rater-by-item combination. There was no case where the percentage of *t*-values was greater or equal to 2. The bias size was quite small (*bias*=-.49). The minimum *t*-value for the interaction was -.04 and the maximum was -1.7.

How the MFRM Complements GT in the Context of this Study

The study evaluated the level of validity of students' evaluation of courses and teaching using GT and MFRM. These models were complementary in explaining the dynamics in the responses provided by the students. Starting with the results from the GT, some level of validity regarding the students' responses was established but not sufficient enough to ensure accurate measurement. This was confirmed by the results of MFRM which revealed that the rating descriptions or scores received by the lecturers did not accurately reflect the actual ability of the lecturers.

The GT results further indicated that the students contributed to the variations in the scores received by the lecturers. The MFRM results gave much insight to the results from GT by highlighting the behaviours of students and the quality of items used for the exercise. For example, GT found that students systematically differed on how they rated their lecturers. The MFRM extended this finding by revealing that the students were influenced by other factors in rating their lecturer (halo effect) and thus, the majority of the students were

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

lenient in their ratings. The MFRM analysis discovered that the inconsistency in the ratings provided by students manifested largely in the form where lecturers received higher ratings than expected. It must also be mentioned that quite a lot of lecturers also received rating which was far lower than their actual ability.

The GT analysis, with respect to the items, only looked at whether the items were homogeneous in terms of difficulty and it ends there. Specific parameters are not computed for individual items to understand how the item functioned. The GT analysis found that item and lecturer-by-item interaction made some contribution to the variability in the scoring, although it was relatively little. The MFRM extends this analysis on items to include descriptions on (1) the clarity of the item structure and soundness of the item, (2) appropriateness of items factored under each of the criteria, and (3) the scale functioning of the items based on how it was used by students. The MFRM analysis found that the majority of the items were problematic. It is either the items were redundant or unclear or did not psychometrically and accurately reflect the criteria (sub-scale) under which they were placed. Further results from the MFRM found non-functional rating scales such that the majority of the students were not able to utilise the scale categories as expected. Just as was found in the GT analysis that students-by-item interaction contributed largely to the variations in the scores received by the lecturers, so was it revealed in the MFRM analysis that the raters were not able to appropriately utilise the rating scales for the items.

Another significant aspect of the GT analysis is the optimisation conducted to provide information on how many evaluation items and students

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

are needed in obtaining an optimum level of dependability in students' appraisal of courses and teaching. Indeed, the results provided how many items and students are needed to achieve a sufficient level of validity in the evaluation exercise. The MFRM analysis provides caution on the measurement of the construct. That is, irrespective of the number of items and students needed, it must be noted that conditions must be right. These right conditions should include training raters (students) to accurately rate and utilising items and scales which ensure that the traits being measured are precisely measured if possible or errors are reduced to the barest minimum.

Discussion

How non-response influences the validity of students' appraisal of courses and teaching

The study discovered that non-response was common among students during the appraisal of courses and teaching in the university. That is, the majority of the students were found not to either participate in the evaluation exercise or did not respond to several items. This result appears not so surprising since several pieces of literature have indicated that non-response is a common feature of online survey (Reisenwitz, 2015; Goos & Salomons, 2017; Guder & Malliaris, 2013). This notwithstanding, the issue of high non-response in this context cannot entirely be attributed to the transition of paper-and-pencil survey to online survey. This is because data on non-response during the paper-and-pencil evaluation survey from the university was not compared to the non-response from the online survey in this study. The high rate of non-response can also be attributed to the non-availability of the internet, high cost of internet data in low-and-middle-income countries that make internet usage prohibitive

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

especially on issues like course evaluation for which students might not see direct benefits. Other determinants include the high number of items on the evaluation form, low interest or motivation to participate in the evaluation exercise, absence of technology gadget for use (e.g., internet supported phones, laptops, etc). Despite these factors, one thing is, however, clear that online survey is associated with some level of non-response. This is supported by many previous studies conducted in the area of students' evaluation of teaching; these studies found evidence of non-response in online evaluation survey (Adams & Umbach, 2012; Dillman et al., 2002; Sliusarenko, 2013).

This present study further revealed that the frequency of non-responses was negatively related to the accuracy of the measurement of teaching and course quality. That is, a high level of non-response resulted in a high rate of measurement error and a low level of validity of responses provided by students in terms of the appraisal of courses and teaching. This result implied that the more students fail to participate in the evaluation of their instructors, the more the "picture" of the instructors are blurred. It has been found that a unit of non-response results in a random error which affects the quality of data obtained (Dillman et al., 2002) and thus, the least of non-response should not be taken for granted since every survey seeks to generalise the results from the sample to the population (Groves et al., 2009). For example, students who do not respond or respond to very few items may not be taking the appraisal serious and are probably motivated by other factors when participating in appraisal.

The association between the frequency of non-response and accuracy of measurement, as found in this study, have been documented in the literature. A study by Reisenwitz (2015) who examined the non-response bias in student

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

data set. Taking the instructor-by-item interaction, for example, the bias size was very low. The results on the instructor-by-rater-by-item interaction also revealed very little evidence of interactions among the facets. The results showed a non-significant result suggesting that there was no significant interaction effect for instructor-by-rater-by-item combination. From previous literature, the only known research which used MFRM was by Börkan (2017). In Börkan's research, a significant interaction between item-by-course type and item-by-expected grade was found. Unlike this present study, Börkan did not conduct the interaction analysis with the main facet which makes these two studies incomparable.

From the GT analysis, two sources of the variability (i.e., Person-by-item interaction and Items) contributed relatively little to the errors of measurement regarding students' appraisal of courses and teaching. This result was consistent with the results from Quansah (2020), VanLeeuwen et al. (1999), and Li et al. (2018). These scholars conducted GT analysis on students' evaluation data found relatively small variances for items and person-by-item interaction. The results from the MFRM analysis expanded on the results from GT and 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, others were highlighted as redundant in terms of measuring a significant aspect of the trait. Furthermore, the present study discovered that the scale categories used for the rating significantly contributed to the errors of measurement in students' responses on the appraisal of courses and teaching. That is, the students were not able to use the scale categories as expected and this led to

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

non-functional rating scales. Just like the results of this study, Börkan further found that the rating scale (5-points) which was used for the survey did not function as expected.

This present study further discovered that instead of a two-dimensional instrument used, the model identified three psychometric dimensions. Specifically, the course content dimension was identified as unidimensional whereas the mode of delivery section was highlighted as multi-dimensional. It was found that the items were non-equivalent in their level of difficulty. It was easy for the lecturers to obtain a higher score under the mode of delivery dimension than the course content section. The results from Börkan's (2017) study also found little evidence of unidimensionality of the evaluation items which is supporting the results of this present study. Börkan, however, failed to explore this unidimensionality as was done in this study. Eckes (2015) discussed the issue of item structure, rating scale and homogeneity among evaluation items, and argued that these indicators are likely to contribute to rating errors. This idea was supported by the results of this study.

The GT analysis in this research revealed that students (nested in classes) were another major source of variability in students' appraisal of courses and teaching. This result signifies that students systematically differed on 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 the results of this study (e.g., Feistauer & Richter 2016; Quansah, 2020; VanLeeuwen et al., 1999). Feistauer and Richter (2016), for example, found a high level of rater uncertainty variability in students' evaluation of teaching exercise at the

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

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 rater 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 from MFRM showed that the majority of the students were lenient when appraising courses and teaching quality. This could be as a result of the fact that the students were afraid to rate a lecturer poorly Whereas some students were not able to discriminate between different performance levels, others were prone to halo effect- ratings were influenced by other factors other than the specific behaviours to look out for. There are specific instances where students will be caught in the web of 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 may rate the lecturer excellently on his/her teaching quality. In this case, a positive impression created by the lecturer has been carried over by the students when rating the lecturer on a different dimension. In some instances, a student may create a negative impression about a lecturer (based on a specific behaviour exhibited); this can lead to the student providing a low rating for the lecturer. It is not always that the lecturer has created a positive or negative impression for the students, but some raters (students) may just fail to distinguish between what the exact behaviours which are being measured and

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

other behaviours of the lecturer. These behaviours of students which constituted 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.

Just like the results from the GT analysis, high variability was found among raters in the MFRM analysis 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. That is, 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 in an instructor. In another study conducted by Fah and Osman (2011) confirmed that they were a high level of consistency among students in their ratings of instructors. Other studies by Raza and Irfan (2018), and Samian and

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Noor (2012) had results that supported the presence of homogeneity of ratings among raters. The results from these studies which employed CMT are not comprehensive as they focus on only one error. The results of these studies only operate on the assumption of stability of student ratings over time neglecting error 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' rating of instructors.

Level of dependability of students' appraisal of courses and teaching

The findings from the GT analysis in this study found that the extent of dependability of students' appraisal of courses was low compared to the dependability of ratings by students on the evaluation of teaching. Thus, the dependability of data regarding the appraisal of their lecturers and courses was relatively moderate and the measure of lecturer proficiency was inaccurate. This result was also confirmed by the results from the MFRM analysis which revealed that lecturers' ability (i.e., teaching proficiency) were not accurately represented by the ratings of the students. The majority of the lecturers obtained scores that were inconsistent with their actual performance. Whereas some lecturers obtained low scores which were supposed to be high, the majority of the students provided high scores which were far above the expected proficiency score of the instructors. It was also revealed that the lecturers were not equal in terms of their teaching ability which was not possible if accurate ratings were provided.

The results regarding the variability of students' appraisal of courses and teaching cast doubts on the level of validity of the responses provided by the students. Students leniency or inconsistent rating, the presence of halo effect,

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

non-functional evaluation items and rating scale, high item variability, among other factors make a judgement about the validity of student ratings of teaching a negative one. This may be so because the evaluation form has not been well-structured and validated and raters are not adequately trained or briefed on how to respond to the evaluation form.

Results from previous studies on the validity of students' evaluation of courses and teaching are mixed. The results of several empirical studies are found consistent with the results of this study; these studies have revealed that the responses of students concerning the evaluation of their instructors are not accurate enough (Börkan, 2017; Braga et al., 2011; Feistauer & Richter, 2016; Goos & Salomons, 2017; Li et al., 2018; Quansah, 2020; Sliusarenko, 2013). A more interesting study for comparison is that study by Quansah which was the only research conducted in Ghana on this area. The results from Quansah's study supported the outcome of this research, which may be as a result of the fact that the study was conducted in a similar setting. Surprisingly, most of the study which had consistent findings either adopted GT or MFRM in their methodology; this could also explain why the results were similar.

Results from other studies were in contradiction to the results of this study; these studies have found that students' responses to the evaluation of their instructors had a relatively high validity (Fah & Osman, 2011; Raza & Irfan, 2018; Samian & Noor, 2012; VanLeeuwen et al., 1999; Zhang et al., 2015). This inconsistency may be a result of the differences in the methodology adopted and the context of the studies. Whereas this study combined both GT and MFRM, the majority of the studies which had contradictory results, compared to this study employed the CMT approach.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

The number of items and raters needed to obtain an optimum level of dependability in students' appraisal of courses and teaching

The results from GT analysis indicated that 18-items and 80 students are needed to achieve optimum dependability in students' appraisal of courses and teaching for large classes. For small-sized classes between 1-25, it was difficult achieving optimum dependability in appraising courses even when the items were up to 30. The dependability index for classes with less than 20 students was very low. The study further revealed that an optimum level of dependability of responses on teaching appraisal would be achieved with12-items and 20 students. This level of dependability is even higher when there are 24-items with 80 students. For small-sized classes with less than 13, it was difficult achieving a sufficient dependability level of responses on teaching evaluation. Small-sized classes between 13 and 25 would, however, produce accurate responses in the measurement of teaching quality.

It must be mentioned that none of the previous studies which used GT analysis did a comprehensive analysis on the optimisation as compared to this study. Out of the four studies which used GT, two conducted optimisations whereas the rest did not. For those studies which conducted optimisation, a relatively low number of items and/or raters were indicated as facets needed to establish good dependability. In a study by Li et al. (2018), which was conducted in universities in China, it was recommended that 20 student raters were needed to obtain a sufficient level of dependability. In another study by Quansah (2020) conducted in a study in Ghana, it was recommended that a single occasion, 20-items, and 25 students were sufficient enough to achieve an optimum level of dependability of responses. The results of these studies

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

contradict that of this present study, probably, because of the differences in the number of facets used in these studies.

Chapter Summary

It was discovered that 18-items and 80 students are needed to achieve optimum dependability in students' appraisal of courses and teaching for large classes. For small-sized classes between 1-25, it was difficult achieving optimum dependability in appraising courses even when the items were up to 30. The dependability index for classes with less than 20 students was very low. The study further revealed that an optimum level of dependability of responses on teaching appraisal would be achieved with12-items and 20 students. This level of dependability is even higher when there are 24-items with 80 students. For small-sized classes with less than 13, it was difficult achieving a sufficient dependability level of responses on teaching evaluation. Small-sized classes between 13 and 25 would produce accurate responses in the measurement of teaching quality.

The study found that non-response was prevalent among students during the appraisal of courses and teaching in the university. This non-response was more common among level 200 students and less prevalent among level 600 students. The study further revealed that the frequency of responses was negatively related to the accuracy of the measurement of teaching and course quality. That is, a high level of non-response resulted in a high rate of measurement error and a low level of validity of responses provided by students in terms of the appraisal of courses and teaching.

The study generally found that the majority of the students were lenient when appraising courses and teaching quality. Thus, students were found to be

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

inconsistent in how they rated the lecturers in terms of the course and teaching quality. High variability was found among raters such that the students were not similar in their level of severity when rating. 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, others were highlighted as redundant in terms of measuring a significant aspect of the trait.

It was discovered that instead of a two-dimensional instrument used, the model identified three psychometric dimensions. Specifically, the course content dimension was identified as unidimensional whereas the mode of delivery section was highlighted as multi-dimensional. It was found that the items were non-equivalent in their level of difficulty. It was easy for the lecturers to obtain a higher score under the mode of delivery dimension than the course content section. Further, the results discovered that the scale categories used for the rating significantly contributed to the errors of measurement in students' responses on the appraisal of courses and teaching. That is, the students were not able to use the scale categories as expected and this led to non-functional rating scales.

The study found that lecturers' ability (i.e., teaching proficiency) scores were not accurately represented by the ratings of the students. The majority of the lecturers obtained scores that were inconsistent with their actual performance. Whereas some lecturers obtained low scores which were supposed to be high, the majority of the students provided high scores which were far above their expected proficiency score. It was also revealed that the lecturers

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

were not equal in terms of their teaching ability. The results failed to provide any evidence of a bias interaction in the data provided by the students.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

This chapter present information on three key areas. The chapter starts by providing a summary of the research methods together with the key findings. This is followed by the contribution of the research to knowledge and conclusions, and lastly, recommendations for policy and practice, and suggestions for future studies are outlined.

Overview of the Study

The overarching purpose of this research was to examine the validity of students' evaluation of teaching and courses in the UCC through the lenses of GT and MFRM. The study was grounded in the positivists' paradigm using the hybrid quantitative approach. Two main quantitative designs were used: a two-facets partially nested random unbalanced design and the three-facets design with a partial credit model. The study covered students in UCC. The population was targeted to regular students which were 24,726 under four colleges in the university for the 2019/2020 academic year. The study used secondary data on students from the Directorate of Academic Planning and Quality Assurance, UCC (DAPQA-UCC). All cases and data points were involved in this study. Thus, the sample size was 24,726 regular students. This number of students involved in the study through a census approach.

Ethical clearance was obtained from the Ethical Review Board of UCC after the proposal has been approved. An introductory letter was also be taken

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

from the Head, Department of Education and Psychology. A letter was also obtained from one of the supervisors, together with which the researcher drafted another letter to officially seek permission to access and use the data. Copies of these letters were sent to DAPQA-UCC (specifically addressed to the Acting Director) to formally ask for permission to access and use the data. This was done after there has been an initial contact with the Director of DAPQA-UCC to pre-inform the directorate and discuss with them what the study sought to achieve and the need for the study (Creswell, 2012). Other ethical considerations taken into account were confidentiality, anonymity, and data de-identification. The data were analysed by conducting univariate GT and partial credit MFRM analyses.

Key Findings

The findings of the study are summarised as follows:

1. The study revealed that rater (nested in classes)-by-item interaction contributed the largest error of measurement to the responses of students during the appraisal of courses and teaching. This was followed by the raters (nested in classes) which also contributed significantly to the variability in students' appraisal of courses and teaching. Two sources of the variability (i.e., Person-by-item interaction and Items) contributed relatively little to the errors of measurement regarding students' appraisal of courses and teaching.
2. It was found that the extent of dependability of students' appraisal of courses was low relative to the dependability of ratings by students on the evaluation of teaching. On the whole, the dependability of data

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

regarding the appraisal of their lecturers and courses was relatively moderate and thus, the measure of lecturer proficiency was not accurate.

3. It was discovered that 18-items and 80 students are needed to achieve optimum dependability in students' appraisal of courses and teaching for large classes. For small-sized classes between 1-25, it was difficult achieving optimum dependability in appraising courses even when the items were up to 30. The dependability index for classes with less than 20 students was very low. The study further revealed that an optimum level of dependability of responses on teaching appraisal will be achieved with12-items and 20 students. This level of dependability is even higher when there are 24-items with 80 students. For small-sized classes with less than 13, it was difficult achieving a sufficient dependability level of responses on teaching evaluation. Small-sized classes between 13 and 25 will produce accurate responses in the measurement of teaching quality.
4. The study found that non-response was prevalent among students during the appraisal of courses and teaching in the university. This non-response was more common among level 200 students and less prevalent among level 600 students. The study further revealed that the frequency of responses was negatively related to the accuracy of the measurement of teaching and course quality. That is, a high level of non-response resulted in a high rate of measurement error and a low level of validity of responses provided by students in terms of the appraisal of courses and teaching.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

5. The study found that the majority of the students were lenient when appraising courses and teaching quality. Whereas some students were not able to discriminate between different performance levels, others had ratings which were influenced by other factors instead of the specific construct being measured. Thus, students were found to be inconsistent in how they rated the lecturers in terms of the course and teaching quality. High variability was found among raters such that the students were not similar in their level of severity when rating.
6. 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, others were highlighted as redundant in terms of measuring a significant aspect of the trait. It was discovered that instead of a two-dimensional instrument used, the model identified three psychometric dimensions. Specifically, the course content dimension was identified as unidimensional whereas the mode of delivery section was highlighted as multi-dimensional. It was found that the items were non-equivalent in their level of difficulty. Largely, it was easy for the lecturers to obtain a higher score under the mode of delivery dimension than the course content section. Further, the results discovered that the scale categories used for the rating significantly contributed to the errors of measurement in students' responses on the appraisal of courses and teaching. That is, the students were not able to use the scale categories as expected and this led to non-functional rating scales.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

7. The study found that lecturers' ability (i.e., teaching proficiency) were not accurately represented by the ratings of the students. The majority of the lecturers obtained scores that were inconsistent with their actual performance. Whereas some lecturers obtained low scores which were supposed to be high, the majority of the students provided high scores which were far above the expected proficiency score of the lecturers. It was also revealed that the lecturers were not equal in terms of their teaching ability.
8. The results failed to provide any evidence of a bias interaction in the data provided by the students. Taking the instructor-by-item interaction, for example, the bias size was very low. The results on the instructor-by-rater-by-item interaction also revealed very little evidence of interactions among the facets. The results showed a non-significant result suggesting that there is no significant interaction effect for instructor-by-rater-by-item combination.

Contribution to Knowledge

This research makes four main novel contributions. First, this study is the first of its kind in quality assurance literature to combine two major approaches to investigate the validity of students' appraisal of courses and teaching. This contribution is relevant because the findings of this study have created the awareness that each of the approaches is not sufficient to unearth the dynamics in the evaluation of courses and teaching programme. Each of the models is unique. For instance, the GT mostly focuses on group-specific analysis and takes the object of measurement as a group whereas MFRM studies individual elements in the object of measurements. It is important to emphasise

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

that both group-level analysis and individual case analysis are significant for effective ratings in students' evaluation exercise. Whereas the results from these models complement each other, they also produce results that are specific to each model. The use of the two models is consistent with multiple sources of validity evidence to be gathered, assessed, and combined into a validity argument to support score interpretation and utilisation (Fan & Bond, 2016; Kane, 2012). By this, investigators in the area of quality assurance are placed at an advantageous point to interrogate the accuracy and plausibility of data provided by students (Kane, 2012).

Secondly, this study appears to be the only research in the area of quality assurance that has investigated the quality or functioning of the evaluation scale (using PCM) and how such scales contribute to errors of measurement in their ratings. It must be acknowledged that several studies have validated evaluation forms to reduce errors during the appraisal of courses and teaching (Ogbonnaya, 2019; Sok-Foon et al., 2012; Quansah, 2020). What these studies failed to do is to assess the quality of the scale categories used and how these categories contribute to errors in the appraisal exercise. In a study by Börkan (2017), MFRM was used to understand the variations on students' appraisal of teaching, and the rating scale was found to contribute to the sources of variation in the ratings. However, Börkan's study adopted the rating scale model and thus, how each specific rater used the scale was not explored.

At the local level, the outcome of the study makes a significant contribution to knowledge for the management of UCC in Ghana. This research will draw the attention of the university (UCC) to an area that appears to have received less consideration over the years. In a study in UCC by Kwarteng and

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

colleagues (2014), it was found there were fluctuations in lecturers' overall ratings provided by the students over 5 years (i.e., from 2008/2009 academic year to 2012/2013). In the same year, another research was carried out in UCC by Kwarteng and Doh-fia (2014) to validate the evaluation instrument used for students' appraisal of courses and teaching in the university. Their study found that some of the items were placed under sub-dimensions which were not supposed to be there psychometrically. In 2016, a study on quality assurance was conducted in UCC by Gyimah et al. (2016) which revealed that some lecturers mentioned that students rate instructors who are strict on them. Other lecturers also argued that the students do not have a good value-judgment in assessing their teaching activities (Gyimah et al., 2016).

The researcher acknowledges the transition from the paper-and-pencil form of evaluation to online evaluation, which is a major development in the area. This study has established a high rate of non-response is associated with this transition, but this was not fully tested in this study. This study extends the results of these previous studies by providing insightful information about the rating situation in the university.

Conclusions

The study concluded that the quality of data provided by students in relation to evaluating courses and teaching has validity concerns. The unfair ratings received by lecturers, the non-functional structure of items, less effective functioning of rating scales, high rate of non-response and inconsistent rating behaviours of students cast doubts on the validity of students' appraisal of courses and teaching data. Based on this, the dependability of the student evaluation of courses and teaching can be considered as relatively low.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Especially, when there is high response rate which were led to high random errors.

Recommendations for Policy and Practice

The following recommendations were made for policy and practice based on the outcome of the research and the conclusions drawn:

1. Based on the high non-response rate of the evaluation exercise, DAPQA-UCC should embark on sensitisation and awareness exercise for students on the need to actively take part in the appraisal of courses and teaching in the university to address the issue of non-response. This exercise should aim to reduce non-response and increase the response rate among students. This exercise should be done particularly at orientation sessions for fresh students who are admitted to the university (if not done). This should go beyond making the students aware of the existence of the directorate to enlightening the students on the benefits of being part of the evaluation. For the continuing students, the directorate can assign some staff to various classes to meet them for sensitisation before the evaluation is done. In all these strategies, one thing should be paramount and that is, emphasis should be placed on educating the students to religiously partake in the evaluation exercise and to expose to them the implications associated with not participating.
2. Due to the findings that the appraisal instrument did not provide optimum information for small-sized classes, the Academic Board (UCC) together with DAPQA-UCC should design multiple evaluation forms or instruments for the evaluation of courses and teaching based on the size of the class. It is recommended that for relatively small-sized

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

classes, say less than 10 students, a more qualitative means of appraisal should be used by the university. These strategies may include interviewing all or some of the students or using open-ended questionnaires on the evaluation form for students to respond. For relatively large-sized classes, the traditional closed-ended items can be maintained on the evaluation form for the evaluation exercise.

3. 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, inability to use the rating scales) during the appraisal of courses and teaching. The training should include what (behaviours) they should look out for when appraising, what constitutes excellent, average or high performance, and how to reduce the effect of extraneous variable (such as the friendliness of the lecturer) from influencing their rating. This training should be decentralised to the faculty or departmental level since what constitutes an outstanding mode of teaching may differ from one field of study to another.
4. The study recommends that DAPQA-UCC should the items on the existing evaluation form should be reviewed and further subjected to rigorous validation procedures. This is to make the items more functional in terms of clarity and construct validity. Also, the scale categories used for the rating exercise should be changed or modified by the directorate. Particularly, scale category descriptors should be provided on the evaluation form to guide the ratings of the students. The directorate should engage in regular re-designing and validation of the

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

evaluation instrument in terms of the structure of items, quality and functioning of the scale, adequacy of the number of items, and clarity of the sub-scales utilized. It is suggested that this validation should be conducted at regular intervals for some years to get a well-established instrument for use.

5. DAPQA-UCC should incorporate in their operations the use of MFRM for generating fair scores for lecturers. In such a case, lecturers actual scores received from the students to depict their quality of teaching can be conveniently adjusted by using MFRM and a fair score will be produced. This fair score acts as a true score and can be used for decision making instead of using the observed rating.

Suggestions for Further Studies

Based on the outcome of the research, it was suggested that future studies should conduct GT and MFRM analyses of students' appraisal data in the same university using data from several semesters and not only one semester as in the case of this study. Similar research should also be conducted in other institutions of higher education using the same models to understand what is happening in other universities in terms of the validity of student evaluation data. Future research on students' appraisal of teaching should incorporate other variables such as gender, the programme of study and level of study using GT and MFRM analyses to understand how such variables contribute to measurement errors in the rating exercise.

Further studies should use data which makes it possible to identify the responses of specific student in different course areas; this is to investigate students rating behaviours for different courses they have registered for. Finally,

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

the approach to studying non-response in this study only established that non-response is associated with reliability coefficients and measurement errors. Therefore, it is recommended that future studies should use more in-depth analyses of the trends of non-response to be able to estimate the exact effect non-response has on the dependability of the data set.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

REFERENCES

Abdulghani, H. M., Shaik, S. A., Khamis, N., Al-Drees, A. A., Irshad, M., Khalil, M. S., Alhaqwib, A. I., & Isnani, A. (2014). Research methodology workshops evaluation using the Kirkpatrick's model: Translating theory into practice. *Medical Teacher, 36*(1), 24-29.

Adams, M. J., & Umbach, P. D. (2012). Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education, 53*(5), 576–591.

Algozzine, B., Gretes, J., Flowers, C., Howley, L., Beattie, J., Spooner, F., Mohanty, G., & Bray, M. (2004). Student evaluation of college teaching: a practice in search of principles. *College Teaching, 52*(4), 134-141.

Alharby, E. R. (2006). *A comparison between two scoring methods, holistic vs. analytic using two measurement models, the generalizability theory and the many facet Rasch measurement within the context of performance assessment*. Unpublished doctoral dissertation, The Pennsylvania State University, USA.

Allal, L., & Cardinet, J. (1997). Generalizability theory. In J. P. Keeves (Ed.), *Educational research methodology and measurement an international handbook* (2nd ed., pp. 734-741). Cambridge University Press.

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology, 50*(2), 341-358.

Allison, P. D. (2001). *Missing data: Quantitative applications in the social sciences*. USA: SAGE Publications.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Alter, M., & Reback, R. (2014). True for your school? How changing reputations alter demand for selective U.S. colleges. *Educational Evaluation and Policy Analysis, 36*(3), 346–370.

Ameh, C. A., & van den Broek, N. (2015). Making it happen: Training health-care providers in emergency obstetric and newborn care. *Best Practice & Research Clinical Obstetrics & Gynaecology, 29*(8), 1077-1091.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1974). *Standards for educational and psychological tests.* Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin, 51,* 201–38.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: American Psychological Association.

SAM JONAH LIBRARY
UNIVERSITY OF CAPE COAST
CAPE COAST

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Amoako, I., & Asamoah-Gyimah, K. (2020). Indicators of students' satisfaction of quality education services in some selected universities in Ghana. *South African Journal of Higher Education, 34*(5), 61-72.

Andersen, E. B. (2005). Rating scale model. *Encyclopedia of Social Measurement, 3*, 307–315.

Anderson, H. M., Cain, J., & Bird, E. (2005). Online student course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education, 69*(1), 34-43.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 357-74.

Andrich, D. (1998). Thresholds, steps and rating scale conceptualization. *Rasch Measurement Transactions, 12*, 648–649.

Andrich, D. (2005). Rasch models for ordered response categories. In B. S. Everitt, & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1698–1707). New York, NY: Wiley.

Angrist, J. D., & Guryan, J. (2008). Does teacher testing raise teacher quality? Evidence from state certification requirements. *Economics of Education Review, 27*(5), 483–503.

Arreola, R. A. (2007). *Developing a comprehensive faculty evaluation system: A guide to designing, building, and operating large-scale faculty evaluation systems*. San Francisco, CA: Jossey-Bass.

Arthur, W., Jr., Bennett, W. J., Edens, P. S., & Bell, S. T. (2003a). Effectiveness of training in organizations: a meta-analysis of design and evaluation features. *The Journal of Applied Psychology, 88*, 234–245.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Arthur, W., Tubre, T., Paul, D. S., & Edens, P. S. (2003b). Teaching effectiveness: The relationship between reaction and learning evaluation criteria. *Educational Psychology: An International Journal of Experimental Educational Psychology, 23*(3), 275-285.

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *The Journal of Economic Education, 37*(1), 21–37.

Babcock, P. (2010). Real costs of nominal grade inflation? New evidence from student course evaluations. *Economic Inquiry, 48*(4), 983-996.

Barrett, S. (2005). Raters and examinations. In S. Alagumalai, D. C. Curtis & N. Hungi (Eds.), *Applied Rasch measurement: A book of exemplars – Papers in honour of John P. Keeves* (pp. 159–177). Dordrecht, The Netherlands: Springer.

Batty, A. O. (2015). A comparison of video-and audio-mediated listening tests with Many-Facets Rasch modelling and differential distractor functioning. *Language Testing, 32*(1), 3–20.

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.

Bembenutty, H. (2009). Teaching effectiveness, course evaluation, and academic performance: the role of academic delay of gratification. *Journal of Advanced Academics, 20*(2), 326-355.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Benton, S. L., & Ryalls, K. R. (2016). *Challenging misconceptions about student ratings of instruction (IDEA Paper No. 58)*. Manhattan, KS: The IDEA Center.

Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment and Evaluation in Higher Education, 30,* 593-601.

Beran, T., Violato, C., & Kline, D. (2007). What's the "use" of student ratings of instruction for administrators? One university's experience. *The Canadian Journal of Higher Education, 37*(1), 27-38.

Beran, T., Violato, C., Kline, D., & Frideres, J. (2007). The utility of student ratings of instruction for students, faculty, and administrators: A "consequential validity" study. *The Canadian Journal of Higher Education, 35*(2), 49-70.

Berg, N. (2005). Non-response bias. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 865–873). London, UK: Elsevier.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*, 48–62.

Betoret, F. D. (2007). The influence of students' and teachers' thinking styles on student course satisfaction and on their learning process. *Educational Psychology: An International Journal of Experimental Educational Psychology, 27*(2), 219-234.

Bolliger, D. U. (2004). Key factors for determining student satisfaction in online courses. *International Journal on E-Learning, 3*(1), 61-67.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Bonds-Raacke, J., & Raacke, J. D. (2007). The relationship between physical attractiveness of professors and students' ratings of professor quality. *Journal of Psychiatry, Psychology and Mental Health, 1*(2), 1-7.

Boring, A., Ottoboni, K., & Stark, P. B. (2016). *Student evaluations of teaching (mostly) do not measure teaching effectiveness*. Retrieved from https://doi.org/10.14293/S2 on 23rd September, 2020.

Börkan, B. (2017). Exploring variability sources in student evaluation of teaching via many-facet Rasch model. *Journal of Measurement and Evaluation in Education and Psychology, 8*(1), 15-33.

Borsboom, D., Mellenberg, G. J., & van Herden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Braga, M., Paccagnella, M., & Pellizzari, M. (2011). Evaluating students' evaluations of professors. *The Institute for the Study of Labor, 5620*, 1-54.

Brennan, R, L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*, 1-21.

Brennan, R. L. (2001a). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38*, 285–317.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Brennan, R. L. (2001b). *Generalizability theory*. New York: Springer-Verlag.

Brennan, R. L. (2001c). *Manual for urGENOVA Version 2.1*. Iowa Testing Programs Occasional Paper Number 49. Iowa City, IA: Iowa Testing Programs, University of Iowa.

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice*, *14*, 9–12.

Brownell, J. E., & Swaner, L. E. (2009). High-impact practices: applying the learning outcomes literature to the development of successful campus programs. *Peer Review, 11*(2), 26-30.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York, NY: Routledge - Taylor & Francis Group.

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118*(3), 409–432.

Carroll, E. (2019). *Hybrid research: Combining qualitative and quantitative methods for better results*. Retrieved from https://www.driveresearch.com/market-research-company-blog/hybrid-research-combining-qualitative-and-quantitative-methods-for-better-results/#:~:text=Hybrid%20research%20is%20a%20combination,to%20deliver%20the%20perfect%20solution.&text=Although%20the%20time%20spent%20with,higher%20degree%20of%20data%20reliability. on 23rd September, 2020.

Chang, N., & Chen, L. (2014). Evaluating the learning effectiveness of an online information literacy class based on the Kirkpatrick framework. *Libri, 64*(3), 211-223.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Chen, L. (2016). *Do student characteristics affect course evaluation completion?* A paper presented at the 2016 Annual Conference of the Association for Institutional Research, New Orleans, LA.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593–2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2633–2679.

Chuah, K. L., & Hill, C. (2004). Student evaluation of teacher performance: random pre-destination. *Journal of College Teaching & Learning, 1*(6), 109-114.

Clayson, D. E., Frost, T. F., & Sheffet, M. J. (2006). Grades and the student evaluation of instruction: a test of the reciprocity effect. *Academy of Management: Learning and Education, 5*(1), 52-65.

Coffey, M., & Gibbs, G. (2001). The evaluation of the Student Evaluation of Educational Quality Questionnaire (SEEQ) in UK higher education. *Assessment & Evaluation in Higher Education, 26*(1), 89-93.

Cohen, L., Manion, L., & Morrison, K. (2013). *Research methods in education* (4th ed.). New York, NY: Routledge.

Collins, L. A., Smith, A. J., & Hannon, P. D. (2006). Applying a synergistic learning approach in entrepreneurship education. *Management Learning, 37*(3), 335-354.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches.* Thousand Oaks, CA: Sage Publications.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Creswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson Education, Inc.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Ohio, OH: Cengage Learning Press.

Cronbach, L. J. (1971). Test validation. In R. Thorndike (Eds), *Educational measurement* (pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on the validity argument. In H. Wainer, & H. Braun (Eds), *Test validity*, (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioural measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology, 16*(2), 137–163.

Crooks, T. J., Kane, K. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice, 3*(3), 265–285.

Cunha, J. M., & Miller, T. (2014). Measuring value-added in higher education: Possibilities and limitations in the use of administrative data. *Economics of Education Review, 42*(1), 64–77.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Curran, J. M., & Rosen, D. E. (2006). Student attitudes toward college courses: an examination of influences and intentions. *Journal of Marketing Education, 28*(2), 135-148.

Dillman, D. A., Eltinge, J. L., Groves, R. M., & Little, R. J. A. (2002). Survey nonresponse in design, data collection, and analysis. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 3–26). New York: Wiley.

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method.* San Francisco: Jossey-Bass.

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment and Evaluation in Higher Education, 29*(5), 611–623.

Douglas, D., & Gifford, R. (2001). Evaluation of the physical classroom by students and professors: a lens model approach. *Educational Research, 43*(3), 295-309.

Douglas, J., McClelland, R., & Davies, J. (2008). The development of a conceptual model of student satisfaction with their experience in higher education. *Quality Assurance in Education, 16*(1), 19-35.

Duggan, M., & Carlson-Bancroft, A. (2016). *How Emerson College increased participation rates in course evaluations and NSSE.* A paper presented at the Annual Conference of the Association for Institutional Research, New Orleans, LA.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Dysvik, A., & Martinsen, Ø. L. (2008). The relationship between trainees' evaluation of teaching and trainee performance among Norwegian executive students. *Educational Psychology, 28*, 747–756.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407–424.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analysing and evaluating rater-mediated assessment* (2nd ed.). Frankfurt: Peter Lang GmbH.

Engelhard, G. (2011). Evaluating the bookmark judgments of standard-setting panellists. *Educational and Psychological Measurement, 71*, 909–924.

Esgin, E., & Senkul, P. (2011). *Delta analysis: A hybrid quantitative approach for measuring discrepancies between business process models*. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-21219-2_38

Ewing, A. M. (2012). Estimating the impact of relative expected grade on student evaluations of teachers. *Economics of Education Review, 31*(1), 141–154.

Fah, B. C. Y., & Osman, S. (2011). A case study of student evaluation of teaching in university. *International Education Studies, 4*(1), 44-50.

Fan, J., & Bond, T. (2016). Using MFRM and SEM in the validation of analytic rating scales of an English speaking assessment. In Q. Zhang (Ed.). *Conference proceedings for Pacific Rim Objective Measurement Symposium (PROMS)* (pp. 29-50).

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Feistauer, D., & Richter, T. (2016). How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment and Evaluation in Higher Education, 10*, 1-17.

Felder, R. M., & Brent, R. (2004). How to evaluate teaching. *Chemical Engineering Education, 38*(3), 200-215.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.: pp. 105- 146). New York: Macmillan.

Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: the relations between perceived quality, easiness, and sexiness. *Assessment & Evaluation in Higher Education, 29*(1), 91-108.

Fisher, W. P., Jr. (1992). Reliability statistics. *Rasch Measurement Transactions, 6*(3), 238-242.

Fisk, R. P., Grove, S. J., & John, J. (2014). *Services marketing: An interactive approach* (4th ed.). Mason, OH: Cengage Learning.

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. *Research in Higher Education, 53*(3), 353–374.

Galetzka, M., Verhoeven, J. W. M., & Pruyn, A. T. H. (2006). Service validity and service reliability of search, experience and credence services: a scenario study. *International Journal of Service Industry Management, 17*(3), 271-283.

Gersten, R., & Baker, S. (2002). The relevance of Messick's four faces for understanding the validity of high-stakes assessments. In G. Tondal, &

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

T. M. Haladyna (Eds), *Large-scale assessment for all students*, (pp. 49–66). London: Lawrence Erlbaum Associates.

Goos, M., & Salomons, A. (2017). Measuring teaching quality in higher education: assessing selection bias in course evaluations. *Research in Higher Education, 58*(1), 341-364.

Grammatikopoulos, V., Linardakis, M., Gregoriadis, A., & Oikonomidis, V. (2015). Assessing the Students' Evaluations of Educational Quality (SEEQ) questionnaire in Greek higher education. *Higher Education, 70*(3), 395-408.

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

Greene, B. A., Miller, R. B., Crowson, H. M., Duke, B. L., & Akey, K. L. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. *Contemporary Educational Psychology, 29*(4), 462-482.

Gronau, R. (1974). Wage comparisons—a selectivity bias. *Journal of Political Economy, 82*(6), 1119–1143.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Hoboken, NJ: Wiley.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Gruber, T, Reppel, A., & Voss, R. (2010). Understanding the characteristics of effective professors: the student's perspective. *Journal of Marketing for Higher Education, 20*(2), 175-190.

Guder, F., & Malliaris, M. (2013). Online course evaluations response rates. *American Journal of Business Education, 6*, 333-337.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*(5), 427–39.

Güler, N. (2014). Analysis of open-ended statistics questions with many facet Rasch model. *Eurasian Journal of Educational Research*, 55, 73-90.

Güler, N., Uyanik, K. G., & Teker, G. (2012). *Generalizability theory*. Ankara: Pegem Akademi Publishing.

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. *International Journal of Research in Commerce, IT & Management, 6*(9), 21-26.

Halpern, D. F., & Hakel, M. D. (2003). Applying the science of learning to the university and beyond: Teaching for long-term retention and transfer. *Change, 1*, 2–13.

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher Quality, volume 2 of *Handbook of the economics of education* (pp. 1051–1078), chapter 18. Elsevier.

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality*. (NBER Working Papers 11154) National Bureau of Economic Research, Inc.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology, 94*(3), 638-645.

Heckman, J. J. (1974). Shadow prices, market wages, and labour supply. *Econometrica, 42*(4), 679–694.

Helgesen, O., & Nesset, E. (2007). What accounts for students' loyalty? Some field study evidence. *International Journal of Educational Management, 21*(2), 126-143.

Hill, M. C., & Epps, K. K. (2010). The impact of physical classroom environment on student satisfaction and student evaluation of teaching in the university environment. *Academy of Educational Leadership Journal, 14*(4), 65-79.

Ho, A. D. D., Arendt, S. W., Zheng, T., & Hanisch, K. A. (2016). Exploration of hotel managers' training evaluation practices and perceptions utilizing Kirkpatrick's and Phillips's models. *Journal of Human Resources in Hospitality & Tourism, 15*(2), 184-208.

Holton, E. F. (1996). The flawed four-level evaluation model. *Human Resource Development Quarterly, 7*(1), 5-21.

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

Houston, J. E., & Myford, C. M. (2009). Judges' perception of candidates' organization and communication in relation to oral certification examination ratings. *Academic Medicine, 84*, 1603–1609.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Hu, Y.-L., Hung, C. H., & Ching, G. S. (2015). Student-faculty interaction: mediating between student engagement factors and educational outcome gains. *International Journal of Research Studies in Education, 4*(1), 43-53.

Hucbner, A., & Lucht, M. (2019). Generalizability Theory in R. *Practical Assessment, Research & Evaluation, 24*(5), 1-12.

Hultman, M., & Oghazi, P. (2008). Good looks - good courses: the link between physical attractiveness and perceived performance in higher educational services. In M. Thyne, K. R. Deans, & J.Gnoth, (Eds), *Australian and New Zealand Marketing Academy Conference* (pp. 2588-2597). Dunedin: University of Otago.

Iramaneerat, C., & Yudkowsky, R. (2007). Rater errors in a clinical skills assessment of medical students. *Evaluation and the Health Professions, 30*, 266–283.

Iramaneerat, C., Smith, E. V., & Smith, R. M. (2008). An introduction to Rasch measurement. In J. W. Osborne (Ed.), *Best practices in quantitative methods,* (pp. 50–70). Los Angeles, CA: Sage.

Isely, P., & Singh, H. (2005). Do higher grades lead to favourable student evaluations? *The Journal of Economic Education, 36*(1), 29–42.

Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education, 5*(4), 419–434.

Jol, G., & Stommel, W. (2016). Ethical considerations of secondary data use What about informed consent? *Dutch Journal of Applied Linguistics, 5*(2), 180-195.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125–160.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–35.

Kane, M. T. (2006). Validation. In R. L. Brennan (Eds), *Educational measurement*, (pp. 17–64). Westport, CT: ACE/Praeger.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3-17.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student evaluation of teaching: The generalizability of class means. *Journal of Educational Measurement, 13*(3), 171-183.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*(6), 615–631.

Kherfi, S. (2011). Whose opinion is it anyway? Determinants of participation in student evaluation of teaching. *The Journal of Economic Education, 42*(1), 19–30.

Kim, S. C, & Wilson, M., (2009). A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model. *Journal of Applied Measurement, 10*(4), 408-423.

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs. *Journal of the American Society of Training Directors, 13*, 3–9.

Kirkpatrick, D. L. (1976). Evaluation of training. In R. L. Craig (Ed.), *Training and development handbook: A guide to human resource development* (2nd ed., pp. 301–319). New York: McGraw-Hill.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Lee, M., & Cha, D. (2016). A comparison of Generalizability Theory and Many Facet Rasch measurement in an analysis of mathematics creative problem-solving test. *The Journal of Curriculum and Evaluation, 19*(2), 251-279.

Lewis, K. (2001). Using midsemester student feedback and responding to it. *New Directions for Teaching and Learning, 87*, 33-44.

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

Lidice, A., & Saglam, G. (2013). Using students' evaluations to measure educational quality. *Procedia–Social and Behavioural Sciences, 70*(25), 1009-1015.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1993). Generalizability theory and Many-facet Rasch measurement. In G. Engelhard Jr. & M. Wilson (Eds.). *Objective measurement: Theory into practice* (pp. 85-98). Norwood, NJ: Ablex.

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago: MESA Press.

Linacre, J. M. (1995). Generalizability theory and many facet Rasch measurement, In Wilson, M. Draney, K & Engelhard, Jr. (Eds.). *Objective measurement: Theory into practice*. Norwood, NJ: Abex.

Linacre, J. M. (2001). Generalizability Theory and Rasch Measurement. *Rasch Measurement Transactions, 15*(1), 806-807.

Linacre, J. M. (2003). *A user's guide to FACETS [computer program manual]*. Chicago: MESA Press.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Linacre, J. M. (2006). Demarcating category intervals. *Rasch Measurement Transactions, 19*, 1041–1043.

Linacre, J. M. (2009). *FACETS (Computer program, version 3.66.1)*. Chicago: MESA.

Linacre, J. M. (2010). Transitional categories and usefully disordered thresholds. *Online Educational Research Journal, 1*(3), 1-10.

Linacre, J. M. (2013). *Facets Rasch measurement* (Version 3.71.3) [Computer Software]. Chicago, IL: Winterp.com.

Linacre, J. M., & Wright, B. D. (2002). Understanding Rasch measurement: Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 486-512.

Lubienski, C. (2007). Marketing schools: consumer goods and competitive incentives for consumer information. *Education and Urban Society, 40*(1), 118-141.

Lunz, M., Wright, B., & Linacre, J. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*(4), 331-345.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158–180.

MacNell, L., Driscoll, A., & Hunt, A. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*(4), 291–303.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Mäkinen, J., Olkinuora, E., & Lonka, K. (2004). Students at risk: students' general study orientations and abandoning/prolonging the course of studies. *Higher Education, 48*(2), 173-188.

Marcoulides, G. A. (2000). Generalizability theory. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modelling* (pp. 527–551). San Diego, CA: Academic Press.

Marsh, H. W. (2007). *Students' evaluations of university teaching: A multidimensional perspective*. The scholarship of teaching and learning in higher education: an evidence-based perspective (pp. 319–384). Springer.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Masters, G. N. (2010). The partial credit model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 109–122). New York, NY: Routledge.

McDaniel, C., Jr., & Gates, R. (2012). *Marketing research* (9th ed.). Hoboken, NJ: John Wiley.

McGowen, W., & Osgathorpe, R. (2011). Student and faculty perceptions of effects of midcourse evaluation. *Improving the Academy, 29*, 160-172.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

McNamara, T. F., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555-576.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

McNulty,J., Gruener, J., Chandrasekhar, A., Espiritu, B., Hoyt, A., & Esminger, D. (2010). Are online student evaluations of faculty influenced by the timing of evaluations? *Advances in Physiology Education*, *34*, 213-216.

McPherson, M. A. (2006). Determinants of how students evaluate teachers. *The Journal of Economic Education, 37*(1), 3–20.

McPherson, M. A., & Jewell, R. (2007). Levelling the playing field: Should student evaluation scores be adjusted? *Social Science Quarterly, 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. *Eastern Economic Journal, 35*(1), 37–51.

Mehrens, W. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16–8.

Messick, S. (1989). Validity. In R. L. Linn (Eds), *Educational Measurement* (pp.13–103). New York: American Council on Education, MacMillan.

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6–12.

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analysis. *Behaviour Research Methods, 38*(3), 542-547.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386–422.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*, 189–227.

Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessment of students* (6th ed.). Harlow, England: Pearson Education Limited.

North, B. (2000). *The development of a common framework scale of language proficiency.* New York, NY: Peter Lang Inc.

North, B., & Jones, N. (2009). *Further material on maintaining standards across languages, contexts and administrations by exploiting teacher judgment and IRT scaling.* Strasbourg, France: Council of Europe/Language Policy Division.

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.

Otani, K., Kim, B. J., & Cho, J. I. (2012). Student evaluation of teaching (SET) in higher education: how to use SET more effectively and efficiently in public affairs education. *Journal of Public Affairs Education, 18*(3), 531-544.

Paull, M., Whisted, C., & Girardi, A. (2016). Applying the Kirkpatrick model: Evaluating an Interaction for Learning Framework curriculum intervention. *Issues in Educational Research, 26*(3), 490-509.

Pietersen, C. (2014). Negotiating a shared psychological contract with students. *Mediterranean Journal of Social Sciences, 5*(7), 25-33.

Popham, W. J. (1997). Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9–13.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Porter, S. R., & Whitcomb, M. E. (2005). Non-response in student surveys: The role of demographics, engagement, and personality. *Research in Higher Education, 46*(2), 127–152.

Porter, S., & Umbach, P. (2006). Student survey response rates across institutions: Why do they vary? *Research in Higher Education, 47*(2), 229–247.

Praslova, L. (2010). Adaptation of Kirkpatrick's four-level model of training criteria to assessment of learning outcomes and program evaluation in higher education. *Educational Assessment, Evaluation and Accountability, 22*(3), 215-225.

Prebble, T., Hargreaves, H., Leach, L., Naidoo, K., Suddaby, G., & Zepke, N. (2004). *Impact of student support services and academic development programmes on student outcomes in undergraduate tertiary study: A synthesis of the research.* New Zealand: Ministry of Education

Quaglia, R. J., & Corso, M. J. (2014). *Student voice: The instrument of change.* Thousand Oaks, CA: Corwin

Quansah, F. (2020). An assessment of lecturers' teaching using generalisability theory: A case study of a selected university in Ghana. *South African Journal of Higher Education, 34*(5), 136-150.

Quansah, F., Ankoma-Sey, V. R., & Asamoah, D. (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., Appiah, E., & Ankoma-Sey, V. R. (2019). University students' preparation towards building a knowledge economy in Africa: A case of

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

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,* 224-239.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press.

Raza, S. A. (2014). Measuring the attributes of university teachers as perceived by students: scale development and validation. *TIJ's Research Journal of Social Science & Management-RJSSM, 4*(4), 74-82.

Raza, S. A., & Irfan, M. (2018). Students' evaluation of teacher attributes: Implications for quality in higher education. *Bulletin of Education and Research, 40*(1), 197-214.

Raza, S. A., & Khawaja, F. N. (2013). Faculty development needs as perceived by departmental heads, teachers, and students of Pakistani universities. *Literacy Information and Computer Education Journal (LICEJ), 4*(1), 901-907.

Raza, S. A., Majid, Z., & Zia, A. (2010). Perceptions of Pakistani university students about roles of academics engaged in imparting development skills: implications for faculty development. *Bulletin of Education & Research, 32*(2), 75-91.

Reisenwitz, T. H. (2015). Student evaluation of teaching: An investigation of nonresponse bias in an online context. *Journal of Marketing Education, 31*(1), 1-11.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Rivera, J. C., & Rice, M. L. (2002). A comparison of student outcomes and satisfaction between traditional and web-based course offerings. *Online Journal of Distance Learning Administration, 5*(3), 1-11.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Rubin, D. B., & Little, R. J. A. (1987). *Statistical analysis with missing data*. New York, NY: Wiley.

Saks, A. M., & Haccoun, R. R. (2010). *Managing performance through training and development*. Toronto: Nelson Publication.

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: a decade of progress. *Annual Review of Psychology, 52*, 471–499.

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

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*(6), 922–932.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405–50. Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 13–24.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Simpson, P., & Siguaw, J. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education, 22*,199–213.

Sliusarenko, T. (2013). *Quantitative assessment of course evaluations*. Kgs Lynby: Technical University of Denmark.

Smith, E. V., Jr. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*, 281–311.

Sohr-Preston, S. L., Boswell, S. S., McCaleb, K., & Robertson, D. (2016). Professor gender, age, and 'hotness' in influencing college students' generation and interpretation of professor ratings. *Higher Learning Research Communications, 6*(3), 1-23.

Sok-Foon, Y., Sze-Yin, H. J., & Yin-Fah, B. C. (2012). Student evaluation of lecturer performance among private university students. *Canadian Social Science, 8*(4), 239-243.

Spooren, P., & Van Loon, F. (2012). Who participates (Not)? A non-response analysis on students' evaluations of teaching. *Procedia—Social and Behavioural Sciences, 69*, 990–996.

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

Stapleton, R., & Murkison, G. (2001). Optimizing the fairness of student evaluations: A study of correlations between instructor excellence, study production, learning production, and expected grades. *Journal of Management Education, 25*(3), 269-291.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

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

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. *Accountability in Research, 23*(6), 319-350.

Stobart, G. (2001). The validity of national curriculum assessment. *British Journal of Educational Studies, 49*(1), 26–39.

Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239-261.

Taut, S., & Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learning and Instruction, 46*, 45-60.

Teker, G. T., Güler, N., & Uyanik, G. K. (2015). Comparing the effectiveness of SPSS and EduG using different designs for Generalizability theory. *Educational Sciences: Theory & Practice, 15*(3), 635-645.

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? *New Directions for Institutional Research, 109*, 45–56.

Tripathy, J. P. (2013). Secondary data analysis: Ethical issues and challenges. *Iranian Journal of Public Health, 42*(12), 1478-1479.

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.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

University of Cape Coast, UCC. (2015). *Criteria for appointment and promotion*. Cape Coast: University Press.

Üstünlüoğlu, E., & Can, S. (2012). Student evaluation of teachers: A case study of tertiary level. *International Journal on New Trends in Education and Their Implications, 3*(4), 92-99.

Van Buren, M. E., & Erskine, W. (2002). *The 2002 ASTD state of the industry report*. Alexandria: American Society of Training and Development.

VanLeeuwen, D. M., Dormody, T. J., & Seevers, B. S. (1999). Assessing the reliability of student evaluation of teaching (SET) with generalisability theory. *Journal of Agricultural Education, 40*(4), 1-9.

Vlăsceanu, L., Grünberg, L., & Pârlea, D. (2004). *Quality assurance and accreditation: A glossary of basic terms and definitions.* Bucharest: United Nations Educational, Scientific and Cultural Organization.

Voss, R., Gruber, T., & Szmigin, I. (2007). Service quality in higher education: the role of student expectations. *Journal of Business Research, 60*(9), 949-959.

Wachtel, H. K. (2006). Student evaluation of college teaching effectiveness: a brief review. *Assessment & Evaluation in Higher Education, 23*(2), 191-212.

Wikström, C. (2006). *Classroom assessment and grading – validity issues in the process of selection to higher education*. Paper presented at the NCME-conference, April 8–10, in San Francisco.

Wilhelm, W. B. (2004). The relative influence of published teaching evaluations and other instructor attributes on course choice. *Journal of Marketing Education, 26*(1), 17–30.

 Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Wolbring, T., & Treischl, E. (2015). Selection bias in students' evaluation of teaching. *Research in Higher Education, 1*, 1–21.

Wolming, S. (1999). Validity issues in higher education selection: A Swedish example. *Studies in Educational Evaluation, 25*(4), 335–51.

Wright, R. (1996). A study of the acquisition of verbs of motion by Grade 4/5 early French immersion students. *The Canadian Modern Language Review, 53*(1), 257-280.

Wu, M., & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement, 14*, 339-355.

Yardley, S., & Dornan, T. (2012). Kirkpatrick's levels and education 'evidence'. *Medical Education, 46*(1), 97-106.

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education, 12*(1), 55–76.

Zhang, S., Fike, D., & DeJesus, G. (2015). Qualities university students seek in a teacher. *Journal of Economics and Economic Education Research, 16*(1), 42-54.

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

APPENDICES

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

APPENDIX A

SAMPLE OF EVALUATION FORM

The University of Cape Coast takes a survey every semester to appraise their lecturers and their courses. Every student taking a course in the University, is supposed to complete a survey for all the courses that he/she takes.

Note that the information you provide during this process will be kept strictly **confidential**. Your **name** and **index number** will be kept anonymous. By **honestly** and **candidly** completing this survey, you would be providing this University with valuable feedback for improvement on the quality of education.

Course Outline

Please tick only one of the options as applicable

1. The lecturer made a course outline available to students at the beginning of the course

○ No ○ Yes

Course Content

Please tick only one of the options as applicable

2. The course content was based on the outline provided

○ Less than 70% ○ 70 - 79% ○ 80 - 89% ○ 90% or more

---

3. The course content is likely to be covered at the end of the teaching period.

○ Not likely ○ Slightly likely ○ Likely ○ Very likely

4. The course content was detailed (i.e each week with its specific content, materials to be read, etc)

○ Not detailed ○ Slightly detailed ○ Detailed ○ Very detailed

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Attendance

Please tick only one of the options as applicable

5. The lecturer met the class ...

Less than 70% of the time 70 - 79% of the time 80 - 89% of the time 90% or more of the time

6. The lecturer was most often in class ...

15 minutes or more late 10 -14 minutes late 5 - 9 minutes late On time

7. The lecturer most often left the class ...

More than 15 minutes to time 10 -14 minutes to time 5 - 9 minutes to time At the end of the period

Mode of Delivery

Please tick only one of the options as applicable

8. The lecturer demonstrated knowledge of the subject matter

Not very well Not well Well Very well

9. The lecturer's delivery was well organised and systematic

Not well Not very well Well Very well

10. The lecturer effectively communicated what he/she was trying to teach

Not very well Not well Well Very well

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

11. The lecturer used class time to fully promote teaching and learning

Not very well Not well Well Very well

12. The lecturer encouraged students to read materials beyond the suggested readings and lecture notes

Not very well Not well Well Very well

13. The lecturer encouraged students to analyse issues from different perpectives

Not very well Not well Well Very well

14. The lecturer accepted other points of view

Not very well Not well Well Very well

15. The lecturer made time during lecturers for questions and discussions

Not very well Not well Well Very well

16. The lecturer's responses to questions were appropriate and informative

Not very well Not well Well Very well

17. The lecturer was concerned with the academic progress of students

Not very well Not well Well Very well

Assessment

Please provide a candid response

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

18. Indicate the mode(s) of continuous assessment used in the course

☐ Quizzes ☐ Assignments ☐ Presentation ☐ Term paper

19. The number of assessments given by the lecturer was

○ 0 ○ 1 ○ 2 or more

20. The lecturer graded assessments

○ No ○ Yes

21. The graded assessments were returned in a good time

○ No ○ Yes

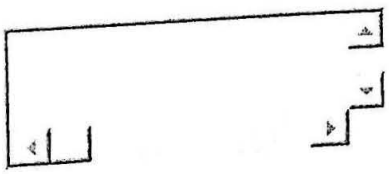
22. The graded assessments were subsequently discussed

○ No ○ Yes

Comments

Please indicate briefly

23. Lecturer's strength(s)

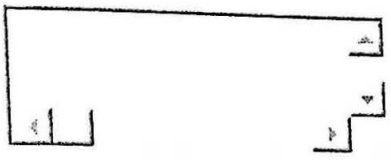


(320 characters remaining)

Digitized by Sam Jonah Library

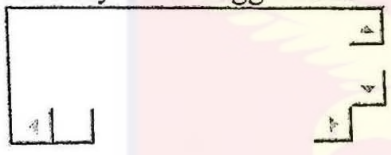
© University of Cape Coast https://ir.ucc.edu.gh/xmlui

24. Lecturer's weakness(es)



(320 characters remaining

25. Any other suggestions



Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

APPENDIX B

INTRODUCTORY LETTER

UNIVERSITY OF CAPE COAST
COLLEGE OF EDUCATION STUDIES
FACULTY OF EDUCATIONAL FOUNDATIONS
DEPARTMENT OF EDUCATION AND PSYCHOLOGY

Telephone: 0332091697
Email:dep@ucc.edu.gh

UNIVERSITY POST OFFICE
CAPE COAST, GHANA

Our Ref

Your Ref

12th February, 2021

TO WHOM IT MAY CONCERN

Dear Sir/Madam,

THESIS WORK
LETTER OF INTRODUCTION
MR. FRANK QUANSAH

We introduce to you Mr. Frank, a student from the University of Cape Coast, Department of Education and Psychology. He is pursuing Doctor of Philosophy Degree in Measurement and Evaluation and he is currently at the thesis stage.

Mr. Frank is researching on the topic, "VALIDITY OF STUDENTS' EVALUATION OF COURSES AND TEACHING USING GENERALIZABILITY THEORY AND MANY-FACET RASCH MODELING IN THE UNIVERSITY OF CAPE COAST."

He has opted to collect or gather data at your institution/establishment for his Thesis work. We would be most grateful if you could provide him the opportunity and assistance for the study. Any information provided would be treated strictly as confidential.

We sincerely appreciate your co-operation and assistance in this direction.

Thank you.

Yours faithfully,

Ms. Ama Atta Ocran
Principal Administrative Assistant
For: HEAD

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

APPENDIX C

ETHICAL CLEARANCE

UNIVERSITY OF CAPE COAST
COLLEGE OF EDUCATION STUDIES
ETHICAL REVIEW BOARD

UNIVERSITY POST OFFICE
CAPE COAST, GHANA

Our Ref: CES-ERB/ucc.ed./vol/21-20
Your Ref:
Date: 10th March, 2021

Dear Sir/Madam,

ETHICAL REQUIREMENTS CLEARANCE FOR RESEARCH STUDY

The bearer, Frank Quansah, Reg. No. ER/MEE/17/0003 an ~~M.Phil.~~ / Ph.D. student in the Department of ~~Education~~ Education and Psychology in the College of Education Studies, University of Cape Coast, Cape Coast, Ghana. He / ~~She~~ wishes to undertake a research study on the topic:

Validity of students' evaluation of courses and teaching through the lenses of generalizability theory and Many-Facet Rasch Modelling in the University of Cape Coast.

The Ethical Review Board (ERB) of the College of Education Studies (CES) has assessed his ~~her~~ proposal and confirm that the proposal satisfies the College's ethical requirements for the conduct of the study.

In view of the above, the researcher has been cleared and given approval to commence his ~~her~~ study. The ERB would be grateful if you would give him ~~her~~ the necessary assistance to facilitate the conduct of the said research.

Thank you.
Yours faithfully,

Prof. Linda Dzama Forde
(Secretary, CES-ERB)

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

APPENDIX D

LETTER FOR DATA ACCESS AND USE

Frank Quansah
Department of Education and Psychology
University of Cape Coast
Cape Coast.

2nd March, 2021.

The Director (Ag)
Directorate of Academic Planning and Quality Assurance
University of Cape Coast
Cape Coast.

Dear Sir,

PERMISSION FOR DATA ACCESS AND USE FOR THESIS WORK

I write to seek for permission to have access to data for my thesis from your directorate on students' evaluation of courses and teaching. I am a PhD student pursuing Measurement and Evaluation with the Department of Education and Psychology. I am identified with the index number EF/MEE/19/0003.

I am writing my thesis which seeks to provide understanding regarding the dynamics of students' appraisal of courses and teaching in the University of Cape Coast using generalizability theory and man-facet Rasch modelling. As part of this research, I will need data from your office on students' evaluation of courses and teaching. This data is necessary to help improve the quality of my problem statement and provide some insights in terms of addressing the research questions.

I will be pleased to have access to a semester students' appraisal of courses and teaching data which have the following indicators: (1) the number of students in the various classes who are required to respond to the evaluation; (2) the number of students in the various classes who are actually responded to the evaluation (including incomplete and complete responses); and (3) the level of study. The data when provided will be used solely for academic purpose and a report on the research work will be given to your office.

I would be glad if I am given the approval to have access for the data and also to permit the use of such data for my thesis. Please find attached a support letter from one of my supervisors. I can be contacted through email: frank.quansah1@stu.ucc.edu.gh or phone: 0541530718. Thank you.

Yours faithfully,



Frank Quansah

Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

APPENDIX E

SUPPORT LETTER FROM SUPERVISOR FOR DATA ACCESS

UNIVERSITY OF CAPE COAST
COLLEGE OF EDUCATION STUDIES
FACULTY OF EDUCATIONAL FOUNDATIONS
DEPARTMENT OF EDUCATION AND PSYCHOLOGY

UNIVERSITY POST OFFICE
CAPE COAST, GHANA

2nd March, 2021

The Director (Ag)
Directorate of Academic Planning and Quality Assurance
University of Cape Coast
Cape Coast

Dear Sir,

LETTER FOR DATA ACCESS AND USE FROM YOUR OFFICE

I write this letter on behalf of Mr. Frank Quansah to seek permission from your office to access and use data on students' appraisal of courses and teaching. I confirm that Mr. Quansah is a PhD candidate in Educational Measurement and Evaluation and currently working on the topic "*Validity of students' evaluation of courses and teaching using Generalizability Theory and Many-Facet Rasch Modelling in the University of Cape Coast Ghana*". I am the co-supervisor.

I would be grateful if Mr. Frank Quansah will be given the approval to have access to the data. The data will help him to effectively carry out his research for his doctoral degree. When given access to the data for use, his findings would make significant contribution in terms of how students' appraisal data in the University of Cape Coast will be understood and utilised. The findings of his research will serve as a prompt for stakeholders in higher education to assess the data provided by students in an attempt to evaluate educational services.

Mr. Quansah is capable of carrying out this research work effectively when granted access to this data. I have no doubt that Mr. Quansah will abide by all ethical issues regarding the use of the data to help university management make informed decisions. I will be grateful if this letter is given the necessary consideration. Thank you.

Yours faithfully,

.....

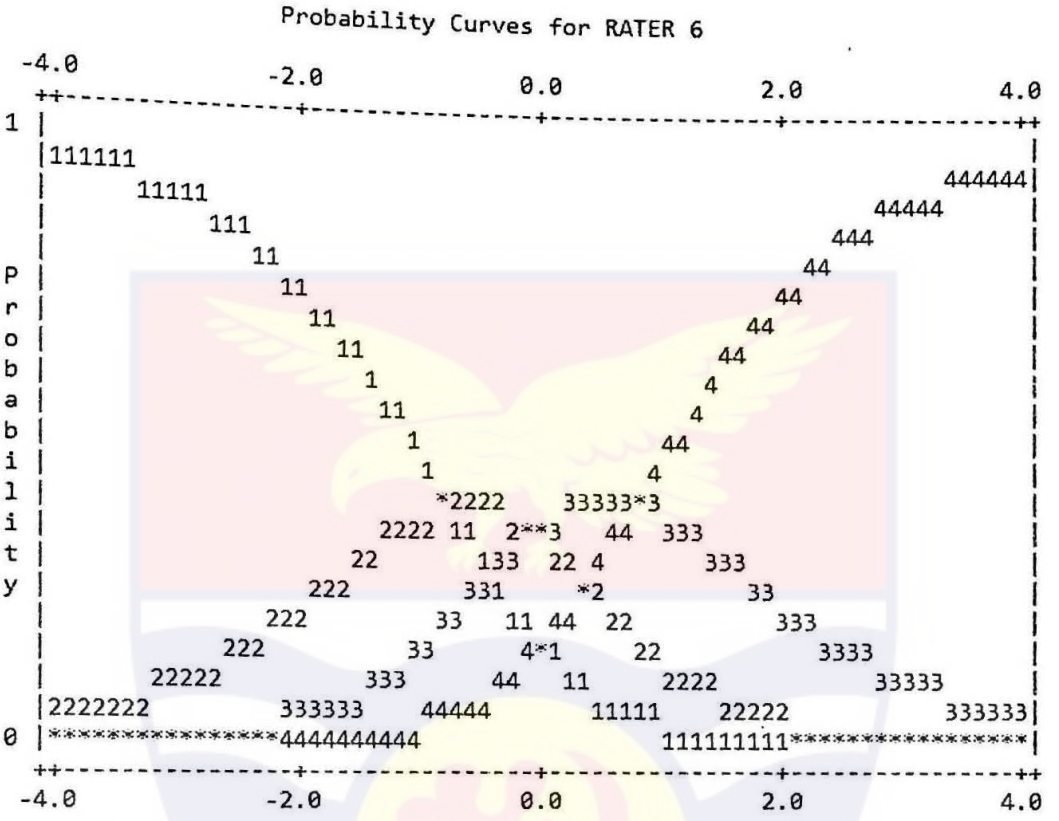
Andrews Cobbinah, PhD
email: andrews.cobbinah@ucc.edu.gh
+233244576092

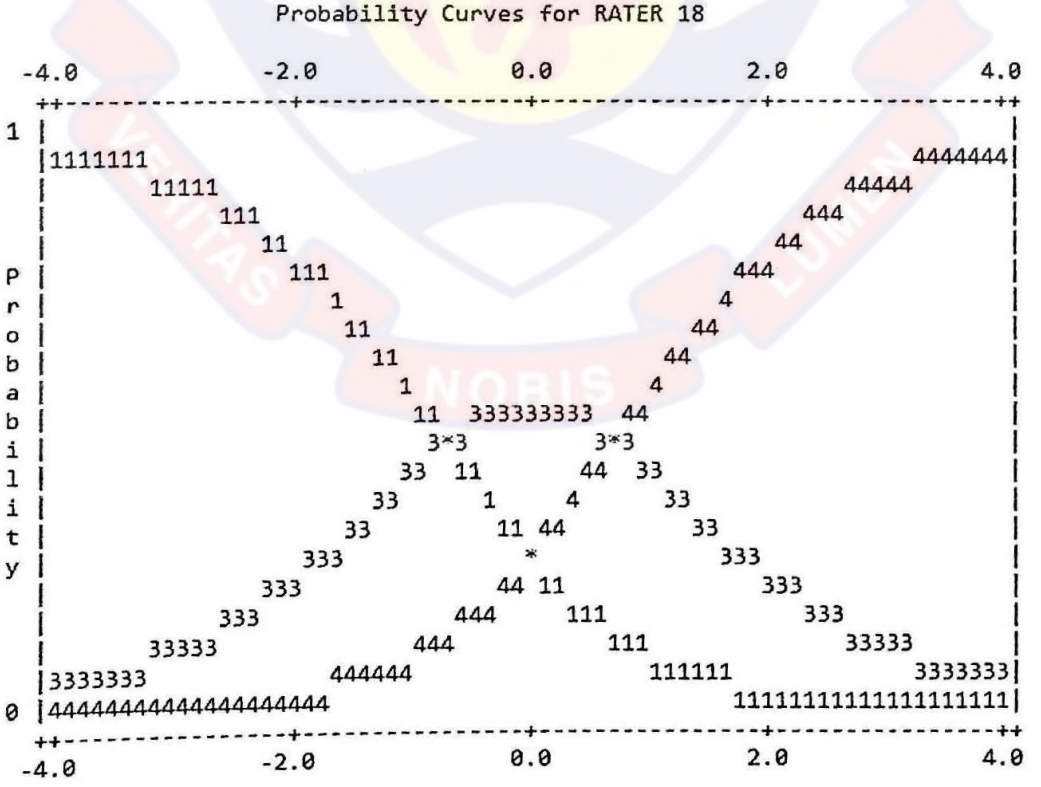
Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

APPENDIX F

SAMPLES OF PROBABILITY CURVES

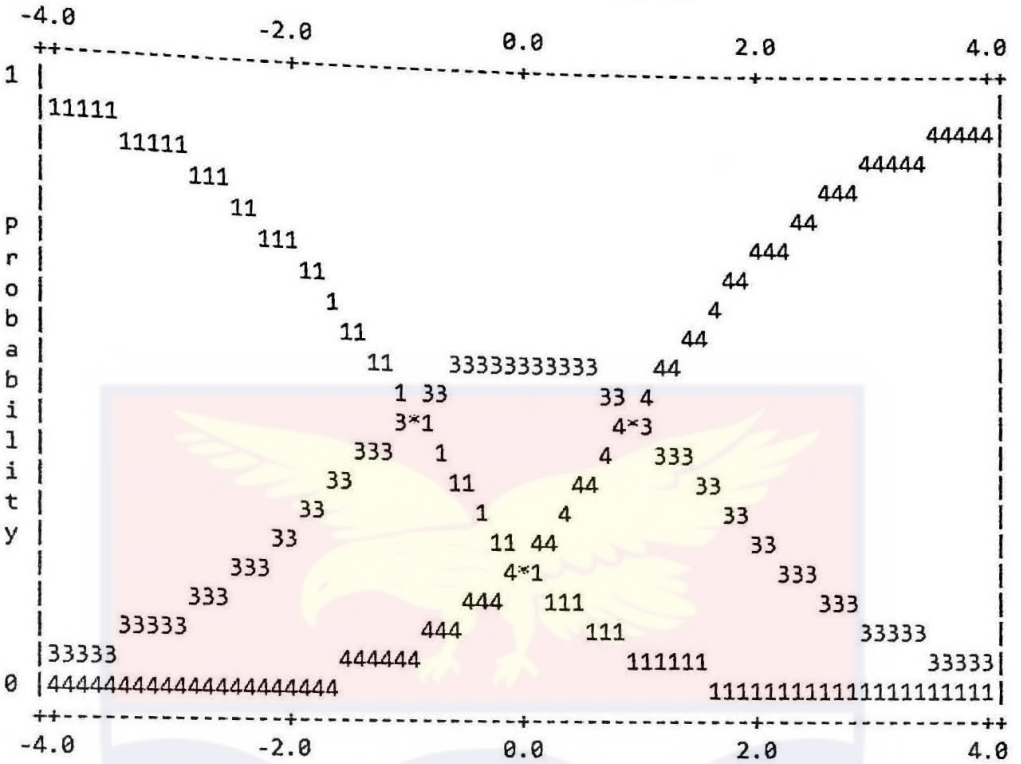




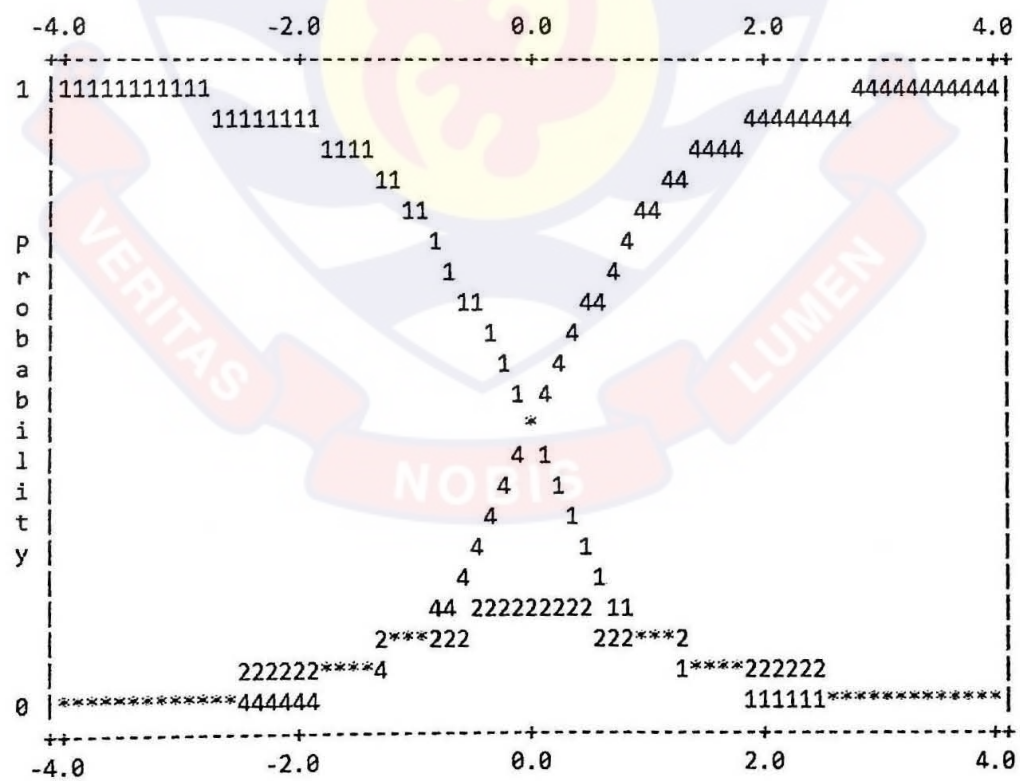
Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui

Probability Curves for RATER 22

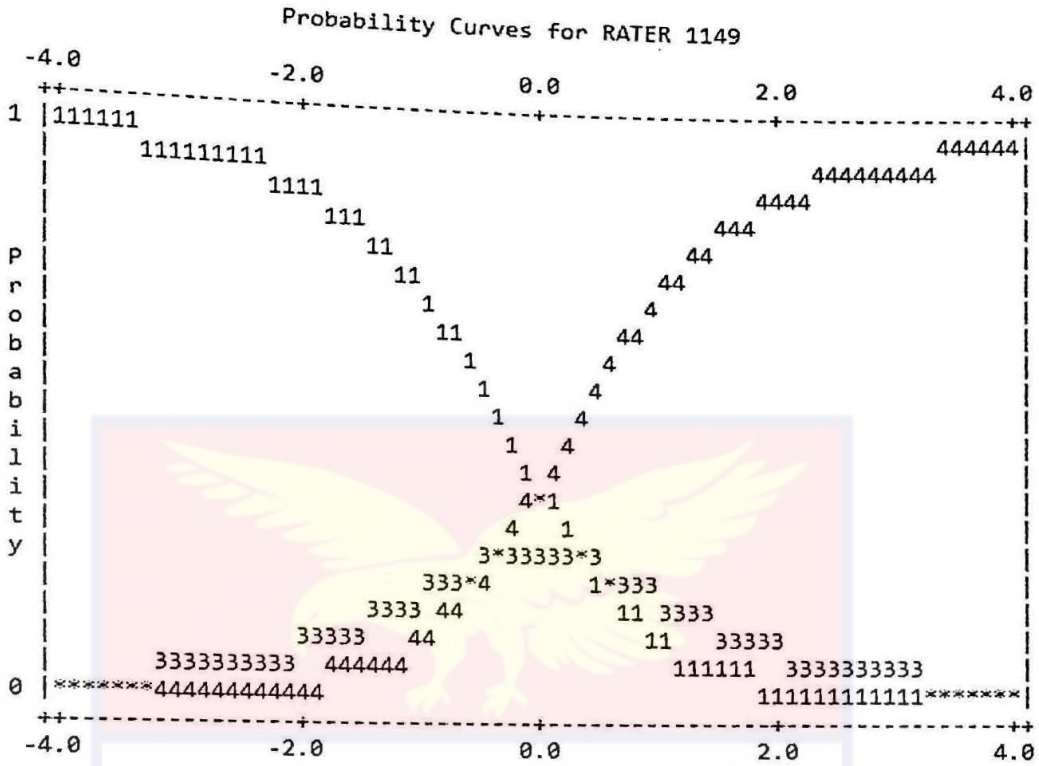


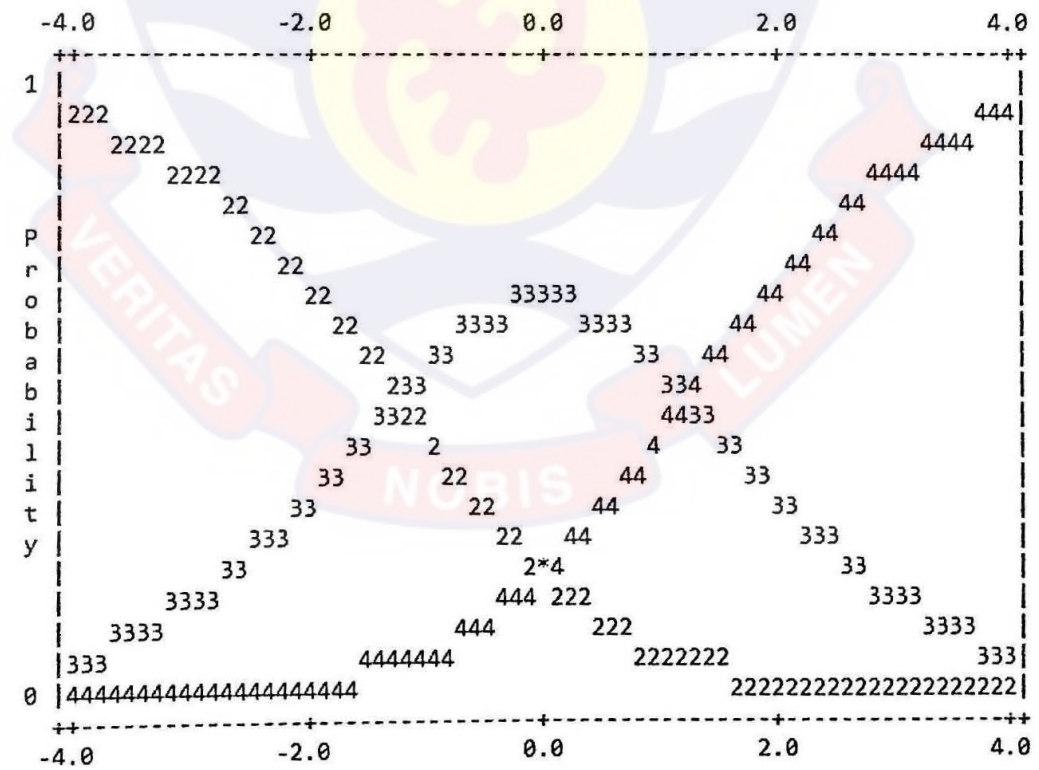
Probability Curves for RATER 19



Digitized by Sam Jonah Library

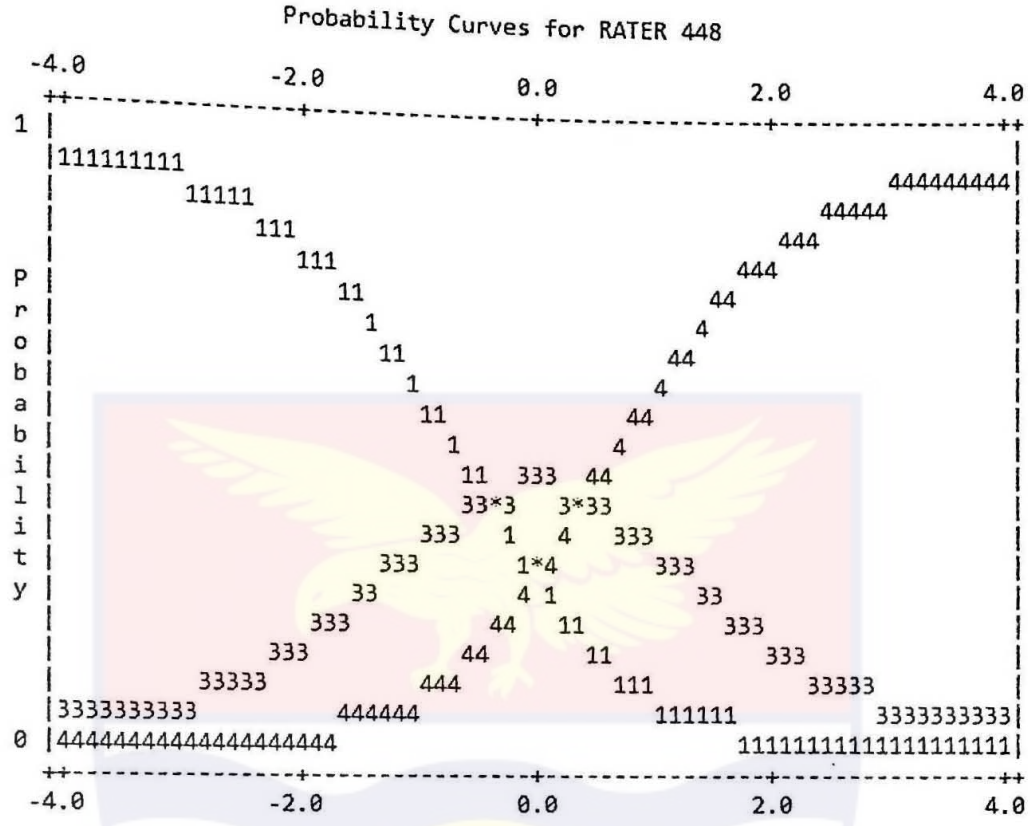
© University of Cape Coast https://ir.ucc.edu.gh/xmlui

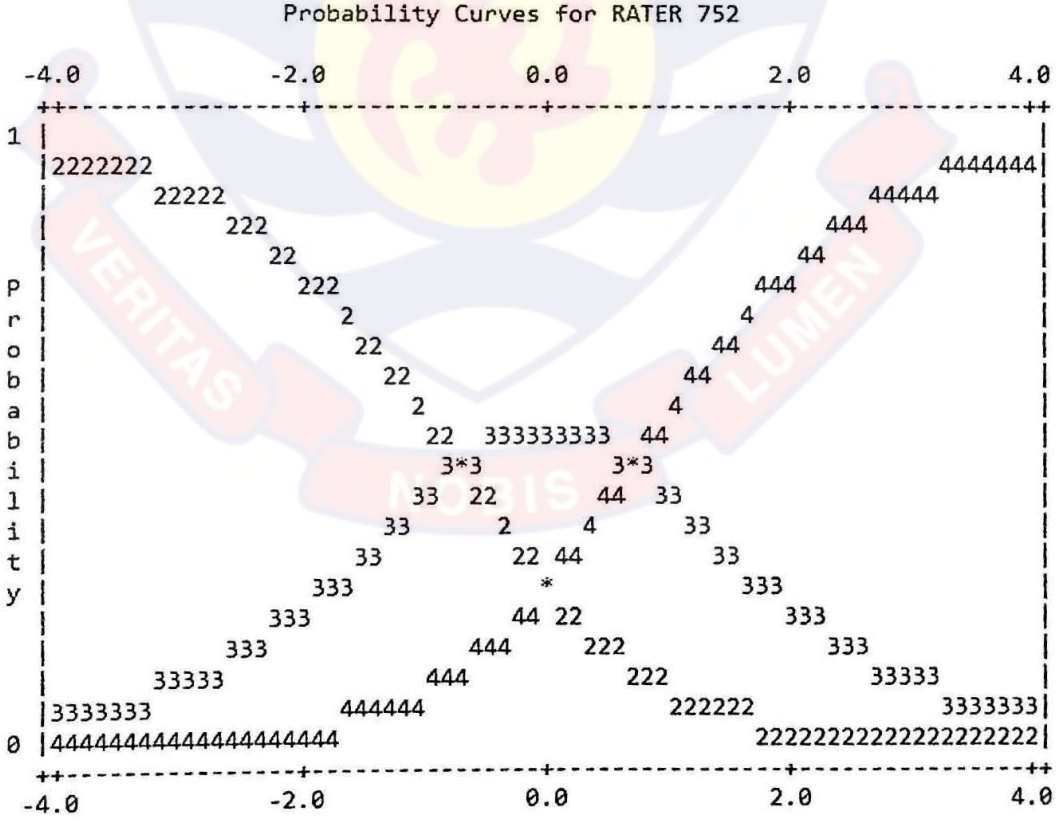




Digitized by Sam Jonah Library

© University of Cape Coast https://ir.ucc.edu.gh/xmlui





SAM JONAH LIBRARY
UNIVERSITY OF CAPE COAST

Digitized by Sam Jonah Library