UNIVERSITY OF CAPE COAST

DEPENDABILITY OF STUDENTS' INTERNSHIP MENTORS' RESULTS USING GENERALIZABILITY THEORY AT UNIVERSITY OF EDUCATION, WINNEBA

GEORGE ODURO-OKYIREH

2020

UNIVERSITY OF CAPE COAST

DEPENDABILITY OF STUDENTS' INTERNSHIP MENTORS' RESULTS USING GENERALIZABILITY THEORY AT UNIVERSITY OF EDUCATION, WINNEBA

BY

GEORGE ODURO-OKYIREH

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

NOVEMBER 2020

DECLARATION

Candidate's Declaration

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

Candidate's Signature:…………………………………………………………….

Date:………………………………………………………………………………….

Name:………………………………………………………………………………...

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:………………………………………………………………………………….

Name:………………………………………………………………………………...

Co-Supervisor's Signature:……………………………………………………..

Date:………………………………………………………………………………….

Name:………………………………………………………………………………...

ABSTRACT

The main purpose of this study was to determine the dependability of the mentors' results of the University of Education, Winneba, School Internship Programme (UEW-SIP), using Generalizability (G) theory. Inherent in this purpose were to find the reliability and the sources of error in the mentors' results. A random effect one-facet crossed design in which intern (p) was crossed with occasion (o), was used for the study. The study used eight purposively selected Faculties out of a total of 14. A total of 9,082 bachelor's degree graduates results for the academic years 2015/2016, 2016/2017 and 2017/2018 were used for the analysis. Data were analysed by performing a univariate generalizability analysis using EduG version 6.1. It was found that for relative interpretation, the results were strongly reliable (Coef_G relative of 0.66 – 0.84) and for absolute interpretation, the results were moderately to strongly dependable (Coef_G absolute of 0.59 – 0.81). The major source of error was the intern by occasion (p x o) interaction. The Intern Teaching Evaluation Form (ITEF), which is the rating scale for evaluating teaching practice in UEW, was found to be reliable to a greater extent. For most dependable mentors' results while ensuring economy of use of resources in the UEW-SIP, the optimum number of occasions was found to be five. It was recommended among other things that, the university supervisors' rating should be increased from one to at least two occasions, so that G theory can be applied to find the psychometric properties of the results. Again, the number of occasions for mentors' rating in the UEW-SIP should be increased from three to five to ensure most stable and dependable results.

KEYWORDS

Generalizability theory

Crossed and nested designs

Universe of admissible observations

Variance components

Classical test theory

Internship programme

ACKNOWLEDGEMENTS

I wish to express my appreciation to all who in diverse ways contributed to the success of this study. My deepest gratitude goes particularly to my principal supervisor Prof. Y. K. A. Etsey, for his fatherly love, encouragement, guidance and support throughout the research and the entire Ph.D. programme. I am greatly indebted to Dr. K. Asamoah-Gyimah, my co-supervisor, for his expert advice and suggestions during this research work. Through his intellectual interaction, a lot have been learnt.

I also express my sincere appreciation to all the lecturers of the Department of Education and Psychology, especially Professor F. K. Amedahe for the diverse ways he contributed to the completion of the Ph.D. programme. My appreciation also extends to my Ph. D. programme mates, Dr. Regina Mawusi Nugba and Dr. Ruth Annan-Brew for their help and encouragement.

Further sincere appreciation goes to the Registrar of UEW, Mr. Paul Osei Barimah (Esq) and the Director of the Institute for Teacher Education and Continuing Professional Development (ITECPD) of UEW, Professor D. K. Danso Mensah for the immense assistance during the data collection stage.

My last appreciation goes to my family and friends for their support, especially, my father, Mr. Mark Oduro Okyireh, my wife, Regina Oduro-Okyireh and brother, Theodore Oduro-Okyireh.

DEDICATION

To my children, Emmanuel, Michelle and Nyamekye.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS/ACRONYMS

CSM	Collaborative School Model
CTT	Classical Test Theory
FAQ’s	Frequently Asked Questions
GES	Ghana Education Service
IEDE	Institute for Educational Development and Extension
IRB	Institutional Review Board
ITECPD	Institute for Teacher Education and Continuing Professional Development
ITEF	Intern Teaching Evaluation Form
NACE	National Association of Colleges and Employers
NGO	Non-Governmental Organisation
NCCBD	Norwegian Center for Child Behavioural Development
OSLC	Oslo and Oregon Social Learning Centre
PMTO	Parent Management Training Oregon
SIP	School Internship Programme
TLM’s	Teaching and Learning Materials
USA	United States of America
UK	United Kingdom
UCC	University of Cape Coast
UDS	University of Development Studies
UEW	University of Education, Winneba
WASSCE	West Africa Senior Secondary Certificate Examination

CHAPTER ONE

INTRODUCTION

In all educational establishments worldwide, that offer professional training programmes in various fields of endeavour, the curricula have two sections that must always be satisfied before graduation and certification. These two sections are the theoretical and practical (internship) aspects of the programme. This plan of training is not different from what pertains to teacher training institutions of which the University of Education, Winneba (UEW), Ghana, is no exception.

The practical aspect of the UEW teacher training programme, which spans the whole of the 7th semester (i.e., 4th year, 1st semester) is evaluated using the Intern Teaching Evaluation Form (ITEF). The evaluation of each student is done by both school-based mentors (three evaluations) and university lecturers (one evaluation). The reliability of the results (scores) of the internship programme has not been looked at since its inception in 2011. The sources of inherent errors which contribute to unreliability of results and the strengths of these inevitable errors are all unknown.

Generalizability theory (G theory), on which this study was based, is able to give cogent answers to the concerns raised above. In addition, an aspect of G theory which is Decision study (D study) was done, leading to a recommendation to redesign the UEW internship programme with respect to the number of occasions of supervision by raters (mentors) required to give the most reliable results and also economise the usage of needed resources

(Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991; Li, Shavelson, Yin & Wiley, 2015).

Background to the Study

It is undeniably impossible for anyone to learn in an educational system and complete at the terminal point and not be exposed to a number of educational and psychological assessment procedures (Oduro-Okyireh, 2013). This is due to the fact that the role of assessment in an educational system is so important that it forms the foundation for nearly all fundamental decision making at all levels of the educational ladder. Continually in an educational system, decisions have to be taken on students, curricula and programmes, and educational policies.

According to Nitko (1996), decisions about students include managing classroom teaching, putting students into different kinds of programmes, assigning them to appropriate class levels, providing guidance and counselling services to them, choosing them for scholastic opportunities and credentialling and certifying their proficiency. Decisions concerning the curriculum and its programmes include decisions on their effectiveness (summative assessment) and on ways to make them better (formative assessment). In Ghana, decisions about education policies are taken at the national level due to the centralised pattern of curriculum development adopted. It must be noted, however, that in the Ghanaian educational system, educational assessments, of which tests and observational strategies dominate, are the major measurement tools that provide almost all the needed information for these types of decisions.

Observational tecniques are used widely in the Ghanaian education system. With practical-oriented courses at the Senior High School (SHS) level such as Agricultural Science, Visual Arts and Home Economics, observational techniques are used in assessing the practical components (Antwi, 1992; Kadingdi, 2006). At the technical and vocational institutions where practical programmes are undertaken, the Technical Education Unit (TEU) of the Ghana Education Service (GES) assesses the practical aspects through internships and written examinations, and observational techniques are also used (TEU, 2010). Again, at the tertiary level in Colleges of Education where teacher training is the core business, the main evaluation tool for the practical aspect is observation (Ministry of Education, 2016). Finally, at the Technical Universities where technical and professional programmes such as Engineering, Accounting, Insurance, Banking and Finance, and so on, are offered, observational techniques are used in the assessment of the practical aspects (National Board for Professional and Technical Examinations [NABPTEX], 2020).

The importance of high standards of education worldwide calls for adequate professional training of the classroom teacher who is the main agent of curriculum implementation. The curriculum for professional teacher training worldwide has two sections which are the theoretical and the practical (internship) aspects. This plan of training pertains to all teacher training institutions in Ghana. Teacher training institutions in Ghana currently are the 46 Colleges of Education (Report of National Conference of Principal of Colleges of Education, PRINCOF, 2018) and four accredited teacher training universities which are University of Cape Coast (UCC), University of Education, Winneba (UEW), University of Development Studies (UDS), Tamale, and Valley View University (VVU), Oyibi, Accra.

The UEW was established in September, 1992 as a University College (University College of Education, Winneba, UCEW) and affiliated to UCC under the Provisional National Defense Council (PNDC) Law 322. The UCEW brought together under one mother institution, seven diploma awarding colleges that were located in different towns in the country. These colleges were the Advanced Teacher Training College (ATTC), the Specialist Training College (STC), and the National Academy of Music (NAM) all located at Winneba; the School of Ghana Languages (SGL) at Ajumako; College of Special Education (CSE) at Akuapem-Mampong; the Advanced Technical Training College (ATTC) at Kumasi; and the St Andrews Agricultural Training College at Mampong-Ashanti. The Winneba Campus was and is still the seat of the Vice-Chancellor and currently with three satellite campuses at Kumasi, Mampong and Ajumako. Each campus is headed by a Principal. On 14th May, 2004, the University of Education Act, Act 672 was passed by the Parliament of Ghana to upgrade the status of the University College of Education, Winneba (UCEW), to the level of a full University and to provide for other related matters. The full university became known as University of Education, Winneba (UEW) (UEW Annual Diary, 2017).

In the 2015/2016 academic year, the total student population was 57,594, which included both under-graduate and post-graduate students. It was made up of 95.2% (54,803) under-graduate and 4.8% (2,791) post-graduate students. It is further broken down into 61% (34,878) male and 39% (22,715) female students. Considering the mode of study, it was 53% (30,525) full-time, 16% (9,215) sandwich, and 31% (17,854) distance students (UEW 21st Congregation Statistics, 2017).

With its mission to “train competent professional teachers for all levels of education as well as to conduct research, disseminate knowledge and contribute to educational policy and development” and vision of “being an internationally reputable institution for teacher education and research,” UEW is tasked with the “responsibility of producing professional educators to spearhead a new national vision of education aimed at redirecting Ghana’s efforts along the path of rapid economic and social development” (UEW Diary, 2017, pp. 3 - 4.). Currently, UEW is the only university in Ghana where about 89.2% of the courses offered have professional educational component (UEW Diary, 2017).

The teacher training programme at UEW is divided into two segments that run alongside each other from Level 100 to Level 400. These two segments are the purely academic courses that prepare the individual in a chosen content area (teaching subjects), and the professional education courses that prepare the individual as a professional teacher. Citing the case of B.Sc. General Agriculture Education Programme offered at the College of Agriculture Education at Mampong-Ashanti Campus, as an example, each academic level for a given semester has one or more core education courses that are taken with other pure academic courses. A student has to undertake a total of between 120 and 144 credit hours to qualify for certification and graduation. Of this number of credit hours required for the four years, the core education (educational foundation) courses take 30 credit hours while methodology and curriculum development courses take three credit hours and teaching practice (on-campus and off-campus) takes seven credit hours (UEW General Agriculture Education Re-accreditation Document, 2013).

The teaching practice programme mentioned in the preceding paragraph is further divided into two segments. These are three credit hours for on-campus teaching practice (peer teaching) under the supervision of two lecturers; and four credit hours for off-campus teaching practice (internship programme) under the mentorship of a Ghana Education Service (GES) staff (mentor) and one UEW Lecturer (Students' Internship Handbook, 2014).

From the inception of the UEW professional educational programmes up to 1998, the practical segment of the programme followed the rational model where the student teacher had to spend more than 90% of the training period on the theoretical aspect of teaching at the student's campus and less than 10% on teaching practice, either on- or off-campus (Students' Internship Handbook, 2014). The main objective of this type of teaching practice was to evaluate the performance of the student to gather scores for graduation requirements without exposing the students to conditions that will help them to develop the right teaching skills and techniques that will help them to become the best of teachers. The worst of this was that the supervision was a one-shot summative activity that was done by only university lecturers without the active involvement of the schools of practice.

To curtail this deficiency in pre-service teacher preparation, in 1999 the UEW restructured its teacher education programme into a four-year Bachelor of Education Degree programme made up of a three-year on-campus segment and a one-year field experience (internship) in partnership with selected schools and colleges. In view of this, at the beginning of the

2002/2003 academic year, UEW established partnerships with a number of schools and colleges to assist in an internship programme in pre-service teacher training. The partnership includes rural, urban and metropolitan schools, both large and small, private and public, geographically spread out throughout the country. The main rationale for such a partnership is expressed in the following words:

> Partnerships for internship programme require schools and colleges to set up quality systems to facilitate the professional growth of student teachers. The role of school-based teacher-mentors has, therefore, become increasingly prominent as they work with their university colleagues in developing systems and procedures to ensure that student teachers are adequately prepared in their teaching career (Students' Internship Handbook, 2014, p.6).

The School Internship Programme (SIP) from 2002/2003 academic year up to 2010/2011 academic year covered a period of one full academic year. It was an extremely thorough school-based student teaching and learning experience that affords a well-thought-out and supervised clinical experience. It did not involve just practicing teaching, but rather undergoing good practices with students in different ways, with the interrogative and thoughtful guidance of a mentor for a full academic year in Basic and Senior High Schools. The UEW-SIP is structured on the Collaborative School Model (CSM) system (Students' Internship Handbook, 2014).

The CSM is a school-university partnership which centres on the development and improvement of classroom teaching and organisational skills for pre-service teachers in diverse settings, through cooperation between mentors, headmasters, proprietors, and university supervisors. To prepare pre-service teachers adequately for the anticipated distinctive challenges of the teaching profession, interns must be engaged in different settings. During those placements, students are given opportunities for exchange of ideas through discussions with experienced and older educationists. Student teachers are also given the opportunity to systematically re-examine their teaching on daily basis in a very supportive environment. Cooperation between the partnership school and university facilitates these important elements and hence emboldens the interns' progress and learning in the development of requisite classroom teaching and organisational skills necessary for use among children from varied ethnic, social, and academic backgrounds" (Dobler, Kesner, Kramer, Resnik & Devin, 2009). The main regulatory principles of the CSM consist of:

i. having interns work together with their peers;

ii. facilitating professional development and improvement opportunities through constant interaction with mentors; and

iii. helping interns to partake in a total school experience (Dobler et al., 2009).

The practice of CSM in school internship programmes ensures that student teachers experience all the current happenings in the teaching profession, thereby practicalising the theoretical knowledge acquired during pre-service training in the light of the current trend of affairs and real-life experiences.

Students in UEW are required to undertake the SIP in Level 400 which is the final year of their studies. From the 2002/2003 academic year, students had to embark on a whole academic year internship in a chosen school for practical training only to come back to the university campus to submit their project work and graduate. In the 2011/2012 academic year, all the Bachelor of Education (B. Ed.) courses at UEW with the exception of three (i.e., B. Ed. Special Education, B. Ed. Basic Education and B. Ed. Early Childhood Education) were changed to Bachelor of Science (B. Sc.) and Bachelor of Arts (B. A.) with education. The rationale for this was to make the UEW academic programmes comparable to the B. Sc. and B. A. programmes undertaken in other universities in Ghana and international universities worldwide. This necessitated the addition of more non-professional education courses to the academic programmes in UEW and this resulted in the internship period being made one semester which is only the first semester of Level 400 instead of a whole academic year (Report of UEW Internship Planning Committee, 2010).

The term internship is explained as a programme where students or professional trainees work in an organisation, which may be with or without pay, so as to acquire work experience and/or satisfy necessary requirements for certification and subsequent graduation (Ohio State University Department of Political Science [OSUDPS], 2013). In this case the internship programme is made an integral part of the professional programme. The student on an internship programme is called an intern.

The UEW-SIP is the type of internship termed as internship for credits (Students’ Internship Handbook, 2014). This means that students undertake this internship compulsorily to obtain credits to satisfy requirements for graduation. The assessment of the intern is a continuous process throughout the internship period by both the school mentor who supervises (rates) at least three lessons of the intern and the university lecturer who supervises (rates) at least one lesson of the intern. From these assessments, a final score is arrived at which enables the intern to be assigned a grade for the internship programme (Students' Internship Handbook, 2014). It is the presence of grades for On-Campus Teaching Practice (Pre-internship) and Off-Campus Teaching Practice (Internship) on a student's academic transcript that certifies partly that the student has gone through a professional teacher training programme and therefore is a professional teacher.

Users of students' teaching practice (internship) results such as the Ghana Education Service (GES) (main employer of teachers in Ghana), private school proprietors, policy makers (Ministry of Education), university administrators and students themselves are undoubtedly indifferent to the particular lessons that were taught and supervised on particular occasions for assessment purposes to meet graduation requirements. They are rather more interested in the performance of the teacher trainee in a number of diverse practical situations in the teaching field as a professional teacher. This is because the greatest mark of every teacher is to exhibit a high sense of professionalism that results in increased academic performance of students. This is the main rationale and substance for teacher recruitment worldwide.

The case above implies that the aforementioned bodies will be greatly willing to substitute other situations of lessons and supervision (rating) conditions for the particular teaching practice experiences that teacher trainees have undertaken and are shown on their academic transcripts. That is, the users of teaching practice results are more interested in the teacher trainees' general teaching practice achievement and more particularly their practical output in the field as teachers than the score they would obtain on any single given occasion of teaching practice. This is pointing towards the teacher's general performance or achievement in teaching as a professional in the field. This generalised achievement is represented by the average score they would have obtained on an infinite number of occasions that they would have been evaluated. This is called the universe score and is the expected value of a person's obtained scores over all observations in the universe of generalisation. This is equivalent to a student's true score in Classical Test Theory (CTT).

To be able to achieve this kind of generalisation, it is pertinent that a student's teaching practice scores obtained from a given number of lessons taught on a given number of occasions with a given number of supervisors (raters) are generalised to estimate the student's average score (universe score) obtainable over all acceptable lessons, occasions and raters.

This study made use of G theory which is a psychometric theory that is grounded on a statistical sampling method that divides observed scores into their underlying manifold sources of variation. Most importantly, it gives a framework that is used to identify and estimate the sources of errors of measurement on which decisions can be made to optimise the measurement procedures so as to give more reliable and dependable scores (Li et al., 2015).

G theory has its basis in the reasoning and argument of Cronbach et al. (as cited in Marcoulides, 2000) that:

> The score on which the decision is to be based is only one of many scores that might serve the same purpose. The decision maker is almost never interested in the response given to the particular stimulus objects or questions, to the particular tester, at the particular moment of testing. Some, at least, of these conditions of measurement could be altered without making the score any less acceptable to the decision maker. That is to say, there is a universe of observations, any of which would have yielded a usable basis for the decision. The ideal datum on which to base the decision would be something like the person's mean score over all acceptable observations, which we shall call his "universe score." The investigator uses the observed score or some function of it as if it were the universe score. That is, he generalises from sample to universe (p. 15).

Cronbach et al. (as cited in Burns, 1998) continue that, the hallmark of G theory is that the accuracy with which the researcher can generalise from a sample of observed scores to the universe score can be estimated prospectively from the information obtained from a G study.

Statement of the Problem

The University of Education, Winneba restructured its teacher education programme into four-year B. Sc. and B. A. with education and B. Ed. Degree programmes which comprised a seven-semester on-campus section and another one-semester field experience (internship) in partnership schools and colleges in the 2011/2012 academic year. The main instrument for supervisors' evaluation of teaching practice is a direct observational schedule, which in effect is a rating scale (Students' Internship Handbook, 2014).

According to Anastasi and Urbina (1987), "direct observations of behaviour play an essential part in personality appraisal, whether in the clinic, counselling centre, classroom, personnel office, or any other context calling for individual evaluations" (p. 463). Evaluation of mentees by supervisors through direct observation is justified by both theoretical assessment principles and practical expediency due to the nature of the attribute under measurement. It would be practically inexpedient to evaluate teaching practice with achievement tests because all achievement tests are evaluations of smaller samples of behaviour exhibited in artificial settings at a given time. Further, these behaviour samples are obtained and evaluated under standardised conditions to be certain of their authenticity and for the assignment of appropriate interpretations. Direct observational techniques for assessment of behaviour (as used in the UEW-SIP) however, provide a more extensive sampling of behaviour in natural settings.

From the onset of the UEW-SIP, there is no literature on the issue of the consistency (reliability) of the measures. In a typical teaching practice programme where students teach different lessons and are evaluated on a number of different occasions using observational techniques by the same or different raters, there would always be issues of measurement inconsistencies across lessons and occasions. Again, the breakdowns of indicators of expected behaviours for the items on the observation schedule (ITEF) (see appendix A) and points to be awarded for such indicators (precise descriptions for scoring) have not been given clearly by the developers. Sometimes, the raters are left to use their own discretions to score lessons as the behaviours unfold. This makes scores suffer from rater subjectivity which may affect reliability.

At the moment, nothing is known about the reliability of the UEW-SIP results for any academic year, the sources of errors in the results, and the magnitudes of these errors that have the potential to culminate into low reliability and validity. There are also no documented facts about the dependability of the UEW-SIP results to aid in decision making, especially concerning teacher recruitment. Knowledge about these psychometric properties of measurement procedures helps in further development and refinement of the measurement procedures. It would also ensure that users of the UEW-SIP results have much confidence in using the results for the intended purposes. Reliable assessment results that are accurate measures of teaching skills would contribute to validity, when interpreted and used appropriately, all things being equal. The gaps in literature pointed out above undoubtedly provide a justification for the conduct of this study.

According to Shavelson and Webb (1991), the CTT which adopts traditional methods of estimation of test reliability, if applied, would only estimate separately, one source of measurement error in one analysis. This for instance, could be, variations in scores across occasions by the test-retest method, internal consistency of items on a test by the split-half and Kuder-Richardson methods or the consistency with which different raters rate the same output (essays) of students by using Cohen’s Kappa which measures inter-rater agreement.

It is in the context of this major limitation of the CTT in reliability estimation that G theory comes in. The power of G theory is that manifold error sources can be identified and separately estimated in one analysis based on which decisions can be made to optimise the measurement procedure used by the supervisors of the UEW-SIP so as to ensure higher consistency of the results (Shavelson & Webb, 1991; Li et al., 2015).

In a summary, put in a question form and partitioned into sub-problems, the main problem of the study is: How reliable are the mentors' results of the UEW-SIP? What are the major sources of error in the results of the UEW-SIP? What alterations can be made concerning the conditions of the measurement design in order to give more dependable results for each academic year? What alterations can be made concerning the conditions of the measurement design in order to optimise the observational technique used in the evaluation of the UEW-SIP? G theory gives the framework for answering the questions raised above and so was applied fully in this study. This is strongly supported by Cronbach et al. (as cited in Lakes & Hoyt, 2009) that researchers developing first-hand measurement procedures should first carry out a G study, which would ultimately be a guide to the design and interpretation of consequent D studies to help come out with the most reliable measurement procedures.

Purpose of the Study

The main purpose of this study was to determine the dependability for decision making, of the mentors' results of the off-campus teaching practice programme in UEW, Ghana, using G theory. Inherent in this overall purpose were a number of specific objectives. These were to determine:

i. the reliability of the mentors' results of the UEW-SIP for each academic year from 2015/2016 to 2017/2018.

ii. the major sources of error in the mentors' results of the UEW-SIP for each academic year from 2015/2016 to 2017/2018.

iii. the number of occasions of rating needed to obtain dependable mentors' results in the UEW-SIP for each academic year from 2015/2016 to 2017/2018 by undertaking D studies which would use the outcomes of already executed G studies.

iv. the number of occasions of rating needed to obtain dependable mentors' results in the UEW-SIP.

Assumptions of the Study

The basic assumptions underlying the use of G theory to estimate reliability and determine dependability are that:

i. the data used in generalizability studies are interval or ordinal by nature. This study used teaching practice scores which are measured with the interval scale;

ii. students' observed scores are composed of their universe scores and one or more sources of error;

iii. the measurement errors are presumed to be independent of the universe score and not correlated with each other. That is, all the effects in the measurement model are independent;

iv. the samples used in estimation of the variance components, which in this study were selected students and occasions, comprise random samples from their individual populations. Nevertheless, the facets can be treated as fixed; and

v. the trait being measured is in a steady state, and any variations among scores obtained by an individual on different occasions of measurement are due to one or more error sources, and not to systematic variations in the person due to cognitive development or learning (Shavelson & Webb, 1991; Hoyt & Melby, 1999).

It must be stated that all the assumptions of the study stated above, were met in the study and none of them was violated. This gives utmost credibility to the findings of the study.

Research Questions

The following four research questions were used to guide the study:

i. How reliable are the mentors' results of the UEW-SIP for each academic year from 2015/2016 to 2017/2018?

ii. What are the identified major sources of error in the mentors' results of the UEW-SIP for each academic year from 2015/2016 to 2017/2018?

iii. What is the optimum number of occasions of rating needed to obtain dependable mentors' results in the UEW-SIP for each academic year from 2015/2016 to 2017/2018?

iv. What is the optimum number of occasions of rating needed to obtain dependable mentors' results in the UEW-SIP?

Significance of the Study

Given the extent of prevalence of the use of observational techniques in evaluation of practical phenomena in natural settings, especially at the tertiary level of education, where professional programmes that have compulsory internship segments are offered, there is the need for research into the reliability and validity of such observational measures. These indicators will likely help users of these results to have confidence in the results.

The findings of the study have pin-pointed the major weaknesses with respect to the sources of error in the mentors' results of the UEW-SIP. This would help the UEW to put appropriate measures in place to either block such errors or at least minimise them. This would likely improve the reliability of the teaching practice results from mentors. This would make the results more dependable for the uses for which they are intended.

In the second place, the findings of the study would provide insight to other Ghanaian teacher training institutions on the need to investigate the reliability of their observational measures. This insight may motivate them to undertake similar studies in their establishments so as to arrive at appropriate psychometric properties to help define institution-wide policy directives to ensure dependable results from their internship programmes.

In the third place, the findings and methods of the study would likely sensitise assessment institutions such as the West Africa Examinations Council (WAEC) and the NABPTEX to do G studies and use the results to do D studies, to enable them to arrive at psychometric indicators that can be used to optimise their measurement procedures. In this case they would be able to design very effective measurement procedures to improve the dependability of their assessment results and also economise the use of resources such as number of items, raters, occasions of rating and time, to ensure cost effectiveness.

Finally, it is hoped that the study would serve as an important reference source for students, universities, colleges of education, polytechnics and other social researchers in the use of G theory in research countrywide. With the results of the study as a point of reference, similar studies can be undertaken to find out the state of affairs with respect to the use of observational techniques as measurement instruments, especially in industrial attachment in business and other professional programmes apart from teaching. This is the first of its kind in the country, so it will undoubtedly open the door for other studies to follow.

Delimitations

The study was confined to UEW first degree regular graduates of the 2015/2016 to 2017/2018 academic years only. These are three consecutive and quite current academic years and so when used for the study would show any current trend in the development of the UEW-SIP with respect to psychometric properties of the results from mentors. This involved all first degree regular graduates in eight faculties with their 35 departments in UEW.

Only the results from the mentors' evaluations in the partnership schools were used for the study with the exclusion of the results from the university supervisors. This is because with the university lecturers' ratings in the UEW-SIP, there was only one rater who rated on only one occasion. Hence, the rater and the occasion facets could not be used since they violated one of the fundamental assumptions of G theory analysis that the number of levels for any facet must be at least two (Crick & Brennan, 1983).

The UEW was used for the study because it is a teacher training university in Ghana with a well-planned internship programme that is an integral part of the requirements for graduation and certification. It is also a teacher training university in Ghana where about 89.2% of the courses offered have a professional educational component. These courses are Bachelor of

Education in specific fields while others are Bachelor of Arts and Bachelor of Science in specific fields with Education.

Finally, G theory was applied fully in this study. The application of G theory for the reliability analysis of the internship results was justified because it offered a framework that was used to achieve the purpose of the study. Application of CTT in reliability studies would have been deficient in achieving this purpose.

Limitations

This study appears to be the first of its kind in social science research in the country and so its execution was met with some challenges which culminated into limitations that affected the outcomes of the study. The main limitations are as follows.

Only the occasion facet was used in the study. This is because the measurement design of the UEW-SIP did not allow the inclusion of the rater facet of the mentors' results and the facets involved in the university lecturers' results in the G theory analysis since these facets contained one level each. This made the estimated variance component of the intern (p) by occasion (o) interaction clouded and difficult to interpret. Thus, a firm conclusion concerning one major source of error in the entire UEW-SIP results could not be made.

The exclusion of the rater facet of the mentors' scores and the facets involved in the university lecturers' scores in the G study analysis, limited the extent of generalisation of the findings of the study. Thus, the findings were generalised to cover only the occasions facet of the mentors' scores and not the entirety of the facets in the mentors' scores and the university lecturers' scores.

The findings of the study could not be generalised beyond UEW to cover other institutions that have off-campus teaching practice internship programmes. This is because the instrument used in rating teaching practice in UEW is context based.

Definition of Terms

Condition: The levels of a facet (e.g., rater 1, rater 2,…,rater k).

Dependability: The correctness of generalising from a person's obtained score on a test or other measures to the average score that person would have received under all the likely conditions that the test user would be equally ready to accept.

Decision study: A study that uses the outcome of a G study to design a measurement procedure that minimises error for a designated purpose.

Dependability Index: Measure of reliability for absolute interpretations.

Facet: A characteristic of a measurement procedure such as item, rater or occasion that is identified as a possible source of measurement error.

Intern: A student on internship.

Internship: A period in an academic programme where students work in organisations to gain professional experience.

G study: A study designed to provide estimates of the variability of possible facets of a measurement procedure.

G coefficient: This is a measure of the accuracy of the generalisation of a person's obtained score to his universe score.

Population: Objects of measurement (usually persons).

Reliability: Consistency of assessment results.

Universe: Conditions of measurement.

Universe of admissible observations: All possible observations that a test user would consider suitable replacements for the observations in hand.

Universe of generalisation: The conditions of a facet to which a decision maker wants to generalise.

Universe score: Symbolised by μ_p, is the expected value of a person's observed scores over all observations in the universe of generalisation. This is same as a person's "true score" in CTT.

Variance component: The variance of an effect in a G study.

Organisation of the Study

The study is organised into five chapters. Chapter one opens with a preamble to the whole study which is followed by the background to the study. Other components of chapter one are problem statement, purpose of the study, assumptions of the study, research questions, significance of the study, delimitations, limitations, and definition of terms. Chapter two centres on the literature related to the study. The literature entails the conceptual framework of the study, the conceptual framework of G theory, and both theoretical and empirical reviews. Chapter three describes the methodology adopted in the study. It examines the research design, the population, the sample and sampling procedure, the research instruments, the data collection method and the data analysis procedure. Chapter four centres on the research results and discussion of the study findings in relation to the reviewed literature. Chapter five gives the summary of the study findings, pertinent conclusions, recommendations based on the research findings and suggestions for further research.

CHAPTER TWO

LITERATURE REVIEW

Introduction

In this chapter, relevant literature to the study was reviewed. This was organised under five broad sections, which are the conceptual framework of the study, conceptual framework of generalizability theory, theoretical framework, school internship programme, and empirical review. The arrangement of the review is as follows.

History of Generalizability Theory

Conceptual Framework of the Study

Conceptual Framework of G theory. This addressed the following concepts:

1. Generalizability Study and"Universe of Admissible Observations"
2. Universe of Generalisation and Decision Study
3. Types of Designs in Generalizability Theory
4. One-Facet and Two-Facet Universes
5. Random and Fixed Facets
6. Variance Components of Crossed Designs with Random Facets
7. Variance Components of Nested Designs with Random Facets
8. Estimation of Variance Components
9. Universe Scores
10. Error Variances
11. Absolute Error Variance, σ^2 (Δ)

12. Relative Error Variance, σ^2 (δ)

13. Coefficients and Indices

The Theoretical Framework. This covered the following thematic areas:

1. Classical Test Theory (CTT)
2. Assumptions of CTT
3. Comparison of Classical Test Theory and Generalizability Theory

The School Internship Programme. This covered the following thematic areas:

1. Types of School Internship Programmes
2. Characteristics of School Internship Programmes
3. Benefits of School Internship Programmes
4. Limitations of School Internship Programmes
5. Components of the UEW School Internship Programme
6. Evaluation of the UEW School Internship Programme
7. The Use of Observational Techniques in Data Collection

The empirical review. This centred on the following subheadings:

1. Generalizability Studies in Ghana
2. Generalizability Studies in the Some Developed Countries

History of G theory

The history about the development of G theory could be traced back to the research findings and publication of a book on measurement theory by Cronbach et al. (1972) entitled "The dependability of behavioural measurements: Theory of generalizability for scores and profiles" (Feldt &

Brennan, 1989; Burns, 1998; Brennan, 1997, 2010). Brennan (2010) then cites Cronbach (1991, pp. 391 – 392) who states that:

> In 1957, I obtained funds from the National Institute of Mental Health to produce, with Gleser's collaboration, a kind of hand-book of measurement theory…. "Since reliability has been studied thoroughly and is now understood," I suggested to the team, "let us devote our first few weeks to outlining that section of the handbook, to get a feel for the undertaking." We learned humility the hard way—the enterprise never got past that topic. Not until 1972 did the book appear … that exhausted our findings on reliability reinterpreted as generalizability. Even then, we did not exhaust the topic. When we tried initially to summarise prominent, seemingly transparent, convincingly argued papers on test reliability, the messages conflicted.

Brennan (2010) continues that, to address these conflicts, Cronbach and his team developed a rich conceptual framework and interwove it with analysis of random effects variance components. The net outcome is "a tapestry that interweaves ideas from at least two dozen authors" (Cronbach, 1991, p. 394, cited in Brennan, 1997, 2010).

It is common to label G theory as the application of ANOVA to CTT. Feldt and Brennan (1989) and Brennan (1997, 2010) argue that this description of the theory is inadequate and misinformative. Rather, it can be correctly suggested that the parentage of G theory can be seen as CTT and ANOVA. G theory (i.e., child), however, is not just the simple combination of its parentage. In actual fact, G theory is not a substitute for CTT, though it liberalises the theory. The statistical mechanism employed in G theory is based on Fisher's (1925) study on factorial designs. But G theory has nothing to do with testing of hypothesis. Instead, it emphasises the evaluation of random effects variance components.

According to Feldt and Brennan (1989), although G theory makes extensive use of ANOVA procedures, application of these procedures to measurement theory and issues started long ago before the work of Cronbach et al. (1972). About three decades earlier before 1972, Burt, Hoyt, Jackson and Ferguson (as cited in Feldt & Brennan, 1989), deliberated on analysis of variance approaches to reliability.

Brennan (1997) notes that for the next several years after the early 1940s, a lot of research on reliability was carried out which marked the framework for G theory. Finlayson's (as cited in Brennan, 1997) study on grades given to essays was undoubtedly the first study of reliability with respect to variance components. Not long thereafter, Pilliner (as cited in Brennan, 1997) brought forth the theoretical relations between intraclass co-relations and ANOVA.

Earlier on, Cronbach (as cited in Brennan, 1997) had indicated the concern that a kind of multi-facet analysis was required to address inconsistencies in reliability estimations. The 1950s marked years when many researchers initiated an investigation into the concept that ANOVA is capable of handling multiple facets concurrently. "Particular examples include Loveland's doctoral dissertation, work by Medley, Mitzel and Doi on classroom observations, and Burt's treatment of test reliability estimated by analysis of variance" (Brennan, 1997, p. 15).

The most outstanding work on multifaceted theory was done by Lindquist (1953). He laid down an extensive highlight of multifaceted theory that concentrated on the estimation of variance components in reliability study. "Lindquist demonstrated that multifacet analyses lead to alternative definitions of error and reliability coefficients. Lindquist's chapter clearly foreshadowed important parts of G theory" (Brennan, 1997, p. 15). This unique contribution by Lindquist is confirmed in the assertion that, "the publications of Burt and Lindquist, in particular appear to have anticipated the development of generalizability theory" (Brennan, 2010, p. 3).

According to Feldt and Brennan (1989) and Brennan (1997, 2006, 2010), on the developmental process of the theoretical framework of G theory, most of the important structures of univariate G theory were completed together with its technical report in 1960 – 1961. This technical report was reviewed in three journal articles with each having a different correspondent author, which were, Cronbach et al., Gleser et al. and Rajaratnam et al. (as cited in Feldt & Brennan,1989; Brennan, 1997, 2006, 2010). It must be noted that, these articles contained the theoretical framework of G theory which was basically, the result of the research which was started in 1957 by Cronbach, Gleser, Nanda, and Rajaratnam with resources from the National Institute of Mental Health to publish a handbook on measurement theory. This feat invariably completed and sealed the work concerning theory advancement of univariate G theory at the time.

According to Brennan (2010), it was not until the mid-1960s that Harinder Nanda carried out a study on the reliability of test batteries that the Cronbach team got motivated to begin the development of the theoretical framework of multivariate G theory (G theory of profiles). This was indeed the principal novel contribution of their work. This was contained in sections of their 1972 monograph (chapters 9 & 10). So, the development of multivariate G theory took another decade.

Brennan (2010) continues that a paper which was published by Cronbach and his team in 1965, largely and accurately, offers a simple but well-designed picture of their early conception of multivariate G theory. In dealing with test batteries, they emphasised the separate treatment of the scores instead of the use of composite scores. This allows the decision maker to study the variances and covariances in the variables and to come out with an optimal D study design. Rajaratnam et al. (as cited in Brennan, 2010) also conceptualised a multivaraite predictor of the universe score. Their paper produced a generalizability theory view about what is now called stratified alpha. It must be noted however, that during the early 1980s, Jarjoura and Brennan (as cited in Brennan, 2010) made an extension of the primary model of multivariate G theory of the Cronbach team and termed it the Table of Specifications Model.

Since the publication of Cronbach et al. (1972), Shavelson and Webb (1981) in their bid to review G theory also discussed further developments in multivariate G theory. The discussions centred primarily on the development of multivariate G coefficient by Joe and Woodward (as cited in Shavelson and Webb, 1981), the clarification of recognised variates in multivariate analyses by Fyans, Salili, Maehr and Desai (as cited in Shavelson and Webb, 1981), and the decision to use either univariate or multivariate analyses for different study designs by Bock (as cited in Shavelson and Webb, 1981).

According to Brennan (1997), since the review of Shavelson and Webb in 1981, there have been other empirical studies and articles by other psychometricians published on multivariate G theory. Brennan (1997) cites examples as Jarjoura and Brennan; Webb, Shavelson and Maddahian; Kolen and Jarjoura; Nupbaum; Brennan; Shavelson, Webb and Rowley; Brennan, Gao and Colton; and Gao, Shavelson, Brennan and Baxter (as cited in Brennan, 1997). All these give perfect and classic illustrations of multivariate analyses in research.

According to Haertel (2006), a major review of the entire theory of generalizability with extensions is provided by Brennan's (2001b) Generalizability Theory. This monograph contains a systematic treatment of multivariate G theory, in which every object of measurement has manifold universe scores, each accompanied with conditions of either one or more fixed facets. The monograph also looks at the estimation of variance components, sampling theory for variance components estimates, conditional standard errors of measurement from a G-theory point of view, and estimation for unbalanced random effects designs. The case of a univariate G theory can best be viewed as a unique case of multivariate G theory.

Conceptual Framework of the Study

Miles and Huberman (1994) have defined a conceptual framework as a pictorial or written product which "explains, either graphically or in narrative form, the main things to be studied—the key factors, concepts, or variables—and the presumed relationships among them" (p. 18). It is mainly an idea or model of what pertains to actuality that a researcher plans to investigate, and the processes involved in these things and the rationale behind—a hypothetical theory of the phenomena that the researcher is studying.

From the purpose of this study, G theory was used to perform a G study first. From the G study, variance components for identified sources of measurement errors were estimated. Generalizability coefficients (G coefficients $[E\rho^2]$) for relative reliability and dependability coefficients (Coef_G absolute $[\Phi]$) for absolute reliability were then computed for the mentors' results of the SIP from 2015/2016 to 2017/2018 academic years. Then, finally, alterations of the conditions of the facet in the measurement (number of occasions) were made in a D study so as to redesign the measurement technique to make it further dependable and economical. A flowchart of the conceptual outline of the study is given in Figure 1.

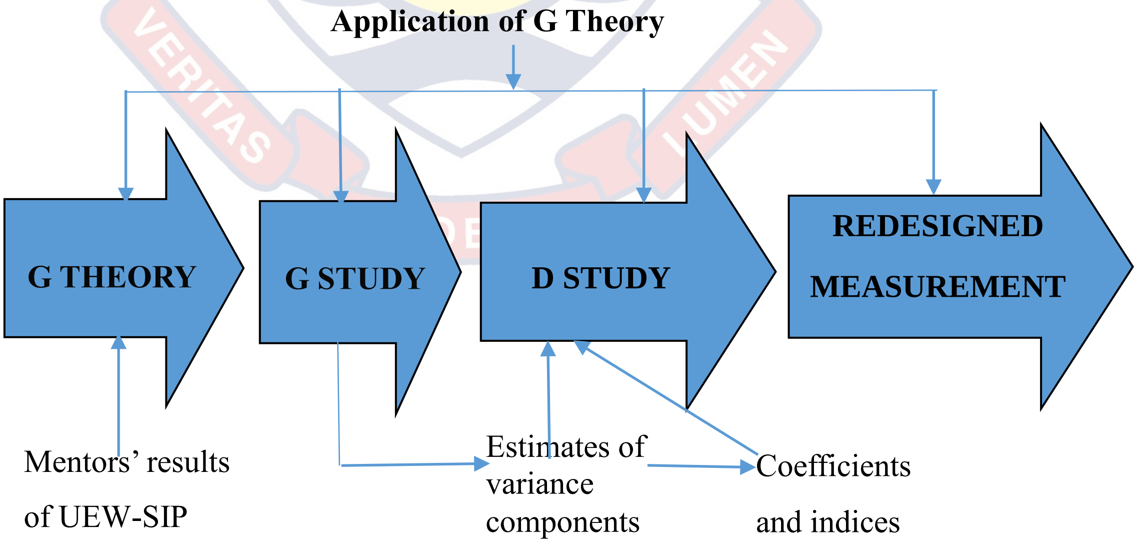


Figure 1: Conceptual framework of application of G theory in the determination of dependability of mentors' results of UEW-SIP

From Figure 1, the thicker arrows show the flow of the study. They indicate the various steps involved in the use of G theory to assess the dependability of scores and the refinement of a measurement procedure. The thinner arrows indicate inputs and products of various steps. First, the results (scores) are fed into the G theory analysis process to perform a designed G study. This produces estimates of variance components for the various variance components identified in the study. Second, the estimates of variance components are used to compute G coefficients, for relative reliability, $E\rho^2$, and absolute reliability, Φ. These are the coefficients and indices. Third, a close study of the estimated variance components and the computed G coefficients is done to enable a D study to be performed by altering the levels of the occasion facet. Fourth, the results of the D study then enable a redesigning of the measurement procedure.

Conceptual Framework of Generalizability Theory (G Theory)

Brennan (2010) asserts that G theory gives a thorough conceptual background and a potent set of statistical techniques for handling numerous measurement problems. On the part of Webb, Shavelson and Haertel (2006), the conception and estimation of reliability by CTT is broadened by G theory. The main basis for the descriptions of G theory above is that by virtue of the provision of a conceptual framework which is grounded in statistics, G theory according to Brennan (2010) allows a researcher to disentangle manifold sources of error that make up the undistinguishable error (*E*) in CTT.

According Brennan (2010), even though CTT and Analysis of Variance (ANOVA) are regarded as the parentage of G theory, G theory is more than just the basic combination of its parentage. To appreciate the inherent value of the concept of G theory demands an understanding that is beyond CTT and ANOVA. The conceptual framework of G theory as it is hinged on CTT and ANOVA with its offshoots from the framework is shown in Figure 2 below (Brennan, 2010).

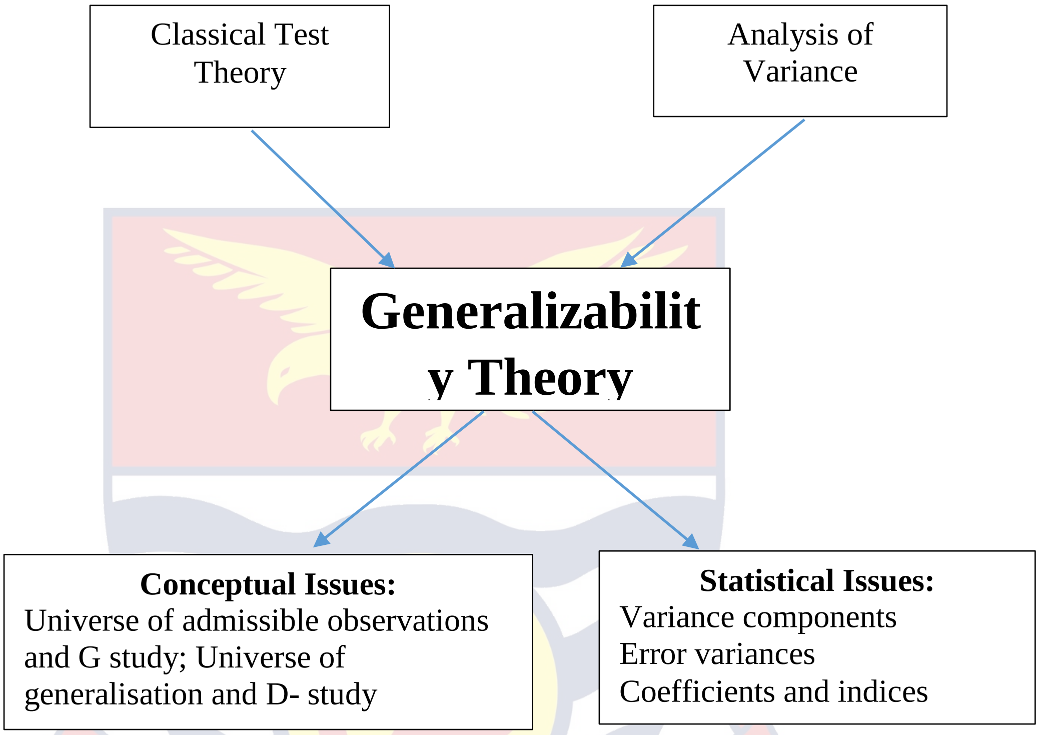


Figure 2: Parents and conceptual framework of generalizability theory

Source: Brennan (2010, p. 5)

It could be seen from Figure 2 that the main concept of G theory is grounded on Classical Test Theory (CTT) and Analysis of Variance (ANOVA), hence, the perfect description that CTT and ANOVA form the lineage of G theory. According to Feldt and Brennan (1989), “G theory can be viewed as an extension and liberalisation of CTT that is achieved primarily through the application of measurement of variance procedures to measurement data” (pp. 127 – 128). The application of measurement of variance procedures is achieved by using factorial ANOVA to divide an individual’s observed score into an effect for the true score and an effect for each source of error, and an effect for each of their combinations (Shavelson & Webb, 1991).

The development of the conceptual framework of G theory which is based on CTT and ANOVA further gives rise to both conceptual issues and statistical issues. The conceptual issues are divided into universe of admissible observations and G-study, and universe of generalisation and D-study while the statistical issues are divided into variance components, error variance, and coefficients and indices.

With the assertion that G theory is more than the simple combination of ANOVA and CTT, Feldt and Brennan (1989), Shavelson and Webb (1991), Haertel (2006) and Brennan (2010), point out that, in CTT, there is no uniform notational system to describe multiplicity of sources of error or alternative definitions of true score and error. An appropriate notational system, critical theoretical distinctions, and related computational techniques are provided by G theory. Regarding the use of ANOVA, Brennan (as cited in Haertel, 2006) simply put it that although G theory uses the statistical machinery of random-effects ANOVA models, it is far more than an application of ANOVA. Rather, G theory should be viewed as a flexible and powerful family of measurement models, encompassing conceptual tools for planning and designing measurement procedures and analysing their error structures.

Generalizability study and universe of admissible observations

According to Shavelson and Webb (1991), "from the perspective of G theory, a measurement is a sample from a universe of admissible observations, observations that a decision maker is willing to treat as interchangeable for the purposes of making a decision" (p. 3). The universe in this context is defined as all the admissible conditions of a given facet of the measurement of interest. “A facet is simply a set of similar conditions of measurement” (Feldt & Brennan, 1989, p. 128). For instance, if the researcher wants to generalise from performance on a given set of achievement test items, say, in Mathematics at the Basic Education Certificate Examination (BECE) level for a given year, to a larger set of mathematics test items at the BECE level in order to make a decision to admit students to the Senior High School (SHS), then ‘items’ is a facet of the achievement measurement. Any set of the test items makes up an admissible condition of measurement for the item facet. The item universe will be defined by the set of all admissible items. That is, the items universe is the set of all other items in Mathematics at the BECE level that the test giver can interchange with those that appeared on the test for the year in question. The test user’s universe of admissible observations contains an item facet which is infinite.

In the UEW-SIP, a decision maker (employer of teachers) could generalise from teaching practice performance on a single occasion to performance on a larger number of occasions. In this case, occasion is a facet. Any one occasion makes up an admissible condition of measurement for the occasion facet. The occasion universe is defined by all admissible occasions (i.e., all occasions of teaching that the teacher trainee can undertake during his professional training). The decision maker’s universe of admissible observations contains an occasion facet which is infinite.

To design a useful and efficient measurement procedure for a designated purpose in the UEW- SIP, an investigator would have to gather and analyse data to experientially describe his universe of admissible observations. To do this, there must be a well-defined number of (sample of n_r) different raters who will evaluate given lessons of a sample of n_p interns on a number of n_o different occasions. This is termed as Generalizability Study (G study).

Universe of generalisation and decision study

The main objective of a G study is to provide as much information as possible about the sources of variation in a given measurement and to obtain estimations of variance components accompanying the universe of admissible observations. A G study defines the universe of admissible observations as largely as possible (Feldt & Brennan, 1989; Shavelson & Webb, 1991; Brennan, 2010).

A Decision study (D study) uses the information provided by the G study on a measurement procedure to redesign the best possible and the most efficient measurement procedure for a designated purpose. According to Shavelson and Webb (1991, p. 12), in planning a D study, the decision maker does the following:

i. describes a universe of generalisation—the number and breadth of facets that he is willing to generalise over;

ii. states the proposed interpretation of the measurement—relative or absolute; the proposed interpretation defines measurement error and thereby identifies the sources of error of greatest concern; and

iii. makes use of the information from the G study about the magnitude of the different sources of measurement error to assess the effectiveness of alternative designs for minimising error and maximising reliability.

A D study chooses only some facets for some designated purpose, and by so doing narrows the score interpretation to a universe of generalisation.

Types of designs in generalizability theory

G theory makes it possible for the decision maker to use different designs in G and D studies with the common ones being crossed and nested designs (Etsey, 2015). In a crossed design, every member of the object of measurement experiences each level of each facet. In a one-facet G study of BECE mathematics test, each student, *p*, responds to each item, *i*. The investigator's universe of admissible observations would be described as crossed. The design of this G study is denoted by $p \times i$, where the symbol, "×", is read "crossed with." It is crossed because given the object of measurement and the item facet, all students respond to all items.

In the UEW-SIP for example, if an investigator is prepared to accept a combination of rater (*r*), occasion (*o*) and person (intern or object of measurement) (*p*), in such a way that all raters will supervise all interns on all occasions, then, the investigator's universe of admissible observations would be described as crossed. It is denoted by $p \times r \times o$. It is crossed because given the rater and the occasion facets, all raters rate all interns on all occasions.

In a nested design, not all levels of one facet experience all levels of another facet. If some conditions of one facet, say, items, *i*, are observed with some conditions of another source of variation, say, persons, *p*, the design of the G study is described as nested. It is denoted by $i{:}p$, where the symbol, ":", is read "nested within." A typical example is the Chemistry practical test at the West Africa Senior Secondary Certificate Examination (WASSCE) level where some students take Alternative A while other take Alternative B of the same test. According to Shavelson and Webb (1991), “facet A is said to be nested within facet B if (a) two or more conditions of A are observed with each condition of B, and (b) different levels of A are associated with each level of B” (p. 55).

In the UEW-SIP example, suppose that each intern is rated by n_r' raters on n_o' occasions. Let the decisions about a person be based on his mean score over the $n_r' \, n_o'$ observations associated with the person. The above is a verbal description of a $p \times R \times O$ crossed design for a D study. There are two main differences between this design and the G study design $p{\times}r{\times}o$. First, the sample sizes for the D study $(n_r' \, and \, n_o')$ need not be the same as the sample sizes for the G study (n_r and n_o). This difference is shown with the use of the primes with D study sample sizes. Second, for the D study, interest is focused on mean scores for persons instead of single person-rater-occasion observations that is focused by G study estimated variance components. The focus on the mean scores is highlighted by the use of the upper-case letters for the raters and occasions facets (Brennan, 1992, 2010).

According to Etsey (2015), G and D studies make use of many designs. The selected design depends on the purpose of the study. Some study designs are, one-facet crossed fixed, one-facet nested random, one-facet nested fixed, one-facet crossed random, two-facet crossed fixed, two-facet nested random, two-facet nested fixed, and two-facet crossed random. The list continues unending. Others are combinations of crossed and nested designs and are called mixed designs.

One-facet and two-facet universes

With the BECE Mathematics test as an example, if item difficulty varies and a person's score depends on the particular sample of items on the test, then generalisation from the sample to the universe is precarious. Item variability therefore represents a possible source of error in generalisation. If items is the only facet being considered here, then the set of admissible items is a single-faceted universe.

With the example of the UEW-SIP, the universe of admissible observations is defined by two facets—raters and occasions. There could be inconsistencies among raters in their ratings of behaviour, so it would be precarious to generalise from one rater's rating of behaviour to the bigger universe of interest. Hence, rater inconsistency indicates a likely source of error in generalisation. Again, with repeated observations of behaviour across occasions, there may be inconsistencies from one occasion to the other. It means that generalisation from the sample of behaviour collected on one occasion to the universe of behaviour across all occasions of interest is precarious. If the rater and occasion facets are the only two facets under consideration, then the set of admissible observations is a two-faceted universe.

Random and fixed facets

It must be noted that when the conditions of each facet can be exchanged for another set of conditions of the same-size from the universe, the estimated variance components pertain to the random-effects model of ANOVA. Shavelson and Webb (1991) explain further that samples representing conditions of a facet are said to be random when the size of the sample is considerably smaller than the size of the universe and the sample is also drawn randomly from the universe or can be exchanged with another sample of the same size drawn from the same universe.

De Finetti (as cited in Webb et al., 2006) adds that, when the conditions of a facet are not sampled randomly from the universe of admissible observations but the proposed universe of generalisation is infinitely large, the notion of exchangeability may be invoked to consider the facet as random. In formalising the notion of randomness of facets, Webb et al. (2006) quoted Feller (1966, p. 225) that, "the random variables $X_1,......,X_n$ are exchangeable if the n! permutations $(X_{k1},...,X_{kn})$ have the same n-dimensional probability distribution. The variables of an infinite sequence X_n are exchangeable if $X_1,.....,X_n$ are exchangeable for each n." An example is that of the BECE Mathematics test, where different sets of items can be used instead of the set of items that were used in a given year and the decision maker would still achieve his intended purpose. This means that conditions in the universe not used in the G study could be exchanged with observed conditions.

On the other hand, when the conditions of each facet is equal to the conditions of the universe to which the decision maker wants to generalise, and exchangeability is not practicable, the facets are treated as fixed facets (Shavelson & Webb, 1991). Webb et al. (2006, p. 24) explain that a fixed facet arises when the:

(a) decision maker purposely selects certain conditions and is not interested in generalising beyond them;

(b) decision maker finds it unreasonable to generalise beyond the conditions observed; and

(c) entire universe of conditions is small and all conditions are included in the measurement design.

An example of a fixed facet is achievement tests that have multiple subtests covering content areas such as Mathematics, Science and Language Skills. Here, all the subjects of interest that form the subtests facet are these three mentioned and so, exchangeability in terms of other subject areas does not make a conceptual sense. The number of conditions of the subtests facet equals the number of conditions in the universe of generalisation.

Variance components of crossed designs with random facets

A one-facet crossed design such as the BECE mathematics test has four sources of variability. One source of variability results from differences among the objects of measurement, which are usually persons. These are differential levels in intelligence, knowledge and skills. The second source of variability results from differential item difficulty. Differential item difficulty would cause generalisation from the item sample to the item universe to be less accurate (Shavelson & Webb, 1991).

The third source of variability results from the educational and experiential background that students bring to the testing situation. For instance, some students may have some relevant previous knowledge to what is being tested while for others the construct being tested becomes a novel experience. The match between a person’s history and a particular item constitutes an interaction between persons and items. This increases variability and causes generalisation from a student’s score on the test sample to his average score over all possible items in the item universe—the universe score, to be less accurate.

The fourth source of variability may result from randomness. Random errors affect individual testees differently but not all testees uniformly. An example is a student who falls sick during testing and so does not perform well. On the other hand, systematic sources of variability could affect a large number of testees uniformly. An example is high temperature or poor ventilation in some testing rooms. All testees in such testing rooms are likely to obtain lower scores than they would have in testing rooms with normal temperature. (Shavelson & Webb, 1991).

Shavelson and Webb (1991) give an outline of the four sources of variability which are called variance components as follows:

i. Differences among objects of measurement.

ii. Differences in item difficulty.

iii. The person-by-item match.

iv. Random or unidentified events.

According to Shavelson and Webb (1991), the third and fourth sources of variability cannot be separated. The reason is that it cannot be exactly known whether after accounting for the first two sources of variability, further differences in scores reflect the person-by-item interaction or random unidentified sources of variability. These two sources of variability are therefore put together as a residual and defined by the person-item interaction ($p \times i$) confounded by other sources of variability denoted by the letter by *e*. The magnitudes of the resultant three types of variation can be estimated by G theory as variance components. They are given in notation in Table 1.

Table 1 - Sources of Variability in One-Facet BECE Mathematics Crossed Measurement

Source of Variability	Type of Variability	Variance Notation
Person (*p*)	Universe score	$\sigma^2{}_p$
Items (*i*)	Conditions	$\sigma^2{}_i$
p×i interaction/unidentified or random	Residual	σ^2pi,e

Source: Adapted from Shavelson and Webb (1991, p. 7)

G theory is used to estimate the magnitudes of the variance components for all the sources of variability as in Table 1. Indices for items and residual, with the exception of the variability from the object of measurement, can be used in a D study to redesign the measurement procedure to make it more reliable.

Considering the case of the 'two-facet universe' design of the UEW-SIP mentioned earlier, the universe of admissible observations is defined by the two facets—raters and occasions. The design of the measurement procedure is crossed and given in notation by $p \times r \times o$. The first source of variability is termed as 'universe-score variability' which is variability attributable to individual differences in the objects of measurement (i.e., differences among the interns). There are other six sources of variability associated with the measurement facets making a total of seven sources of variability which are all variance components (Feldt & Brennan, 1989; Shavelson & Webb, 1991). Table 2 gives the sources of variability.

Table 2 - Sources of Variability in a Two-Facet UEW-SIP Crossed Measurement

Source of Variability	Type of Variability	Variance Notation
Person (*p*)	Universe-score variance (object of measurement)	σ^2_p
Raters (*r*)	Constant effect for all persons due to stringency of raters.	σ^2_r
Occasions (*o*)	Constant effect for all persons due to their behavioural inconsistencies from one occasion to another.	σ^2_o
$p \times r$	Inconsistencies of raters' evaluation of particular person's behaviour.	σ^2_{pr}
$p \times o$	Inconsistencies from one occasion to another in particular person's behaviour.	σ^2_{po}
$r \times o$	Constant effect for all persons due to differences in raters' stringency from one occasion to another.	σ^2_{ro}
$p \times r \times o, e$	Residual consisting of the interaction of p, r, o; unmeasured facets that affect the measurement; and/or random events.	$\sigma^2_{pro,e}$

Source: Adapted from Shavelson and Webb (1991, p. 9)

After these sources of variation have been identified as in Table 2, G theory then estimates the magnitudes of the variance components for all the sources of variability. Indices for raters, occasions, their interactions and residual, with the exclusion of the variability from the object of measurement can be used in a D study to redesign the measurement procedure to make it more efficient.

Measurements in the social sciences are complex in nature and their purposes cannot always be captured by a one-facet or two-facet designs (Shavelson & Webb, 1991). For instance, in the UEW-SIP, where raters and occasions are the two facets already mentioned, if interns plan on different lessons (*l*), from different subject areas to teach on different occasions, then, lessons (*l*), will be the third facet of this measurement. Generalising from performance on one lesson to the average over all possible lessons may lead to error. If the universe of admissible observations is defined in such a way that all acceptable raters (*r*) are assigned to supervise all lessons (*l*) on all given occasions (*o*), then the design of the measurement procedure is crossed and given in notation by $p \times r \times l \times o$. This is a three-facet crossed design.

In analysing data pertaining to this design, the observed score variance is divided into variance due to main effects (i.e., *p*, *r*, *l and o*); two-way interactions (i.e., $p{\times}r$, $p{\times}l$, $p{\times}o$, $r{\times}l$, $r{\times}o$, and $l{\times}o$); three-way interactions (i.e., $p{\times}r{\times}l$, $p{\times}r{\times}o$, $p{\times}l{\times}o$, and $r{\times}l{\times}o$); and the four-way interaction, which is confounded with error ($p \times r \times l \times o, e$) (Lakes & Hoyt, 2009). Hence, there would be 15 sources of variability which include that of the object of measurement. Table 3 gives the sources of variability, the type of variability and variance notations.

Table 3 - Sources of Variability in a Three-Facet UEW-SIP Crossed Measurement

Source of Variability	Type of Variability	Variance Notation
Persons (*p*)	Universe score for person *p* (deviation from grand mean, averaged over raters, lessons, and occasions)	σ^2_p
Raters (*r*)	Rater effect for rater *r* (rater leniency/stringency, averaged over persons, lessons, and occasions)	σ^2_r
Lessons (*l*)	Lesson effect for lesson *l* (deviation from grand mean, averaged over persons, raters, and occasions)	σ^2_l
Occasions (*o*)	Occasion effect for occasion *o* (deviation from grand mean, averaged over persons, raters, and lessons)	σ^2_o
$p \times r$	Peculiar perception of person *p* by rater *r* (averaged over lessons and occasions)	σ^2_{pr}
$p \times l$	Peculiar perception of person *p* on lesson *l* (averaged over raters and occasions)	σ^2_{pl}
$p \times o$	Peculiar perception of person *p* on occasion *o* (averaged over raters and lessons)	σ^2_{po}
$r \times l$	Peculiar leniency of rater *r* on lesson *l* (averaged over persons and occasions)	σ^2_{rl}
$r \times o$	Peculiar leniency of rater *r* on occasion *o* (averaged over persons and lessons)	σ^2_{ro}
$l \times o$	Peculiar effect for lesson *l* on occasion *o*	σ^2_{lo}

Table 3 (Continued)

$p\times r\times l$	Peculiar perception of person *p* by rater *r* on lesson *l* (averaged over occasions)	σ^2_{prl}
$p\times r\times o$	Peculiar perception of person *p* by rater *r* on occasion *o* (averaged over lessons)	σ^2_{pro}
$p\times l\times o$	Peculiar perception of person *p* on lesson *l* on occasion *o* (averaged over raters)	σ^2_{plo}
$r\times l\times o$	Peculiar leniency of rater *r* on lesson *l* on occasion *o* (averaged over persons)	σ^2_{rlo}
$p\times r\times l\times o,e$	Peculiar perception of person *p* by rater *r* on lesson *l* on occasion *o*, confounded with random error	$\sigma^2_{prlo,e}$

Source: Adapted from Lakes and Hoyt (2009, p. 156)

After these sources of variation have been identified as in Table 3, G theory then estimates and interprets the magnitudes of the variance components for all the sources of variability. Indices for raters, lessons, occasions, their interactions, and residual with the exception of the variability from the object of measurement can be used in a D study to redesign the measurement procedure to make it more reliable. It must be noted however that, "the broader the universe of admissible observations, the greater the possibility of making an error in generalising from sample to universe" (Shavelson & Webb, 1991, p. 10).

According to Feldt and Brennan (1989), Shavelson and Webb (1991), Haertel (2006) and Brennan (2010), the components of an observed score for a particular person on a particular item (X_{pi}) in a one-facet crossed design as shown in Table 1 can be decomposed into four as follows:

$$\begin{array}{llllr} X_{pi} & = & \mu & & \text{(grand mean)} \\ & + & \mu_p - \mu & & \text{(person effect)} \\ & + & \mu_i - \mu & & \text{(item effect)} \\ & + & X_{pi} - \mu_p - \mu_i + \mu & & \text{(residual)} \end{array}$$

Shavelson and Webb (1991, p. 19) further explain the above parameters as given below.

i. The grand mean, μ, a constant for all people, positions the score on the particular scale of measurement.

ii. The person effect shows the distance between an individual's universe score (μ_p) and the grand mean (μ). A positive person effect indicates that the person scored higher than average; a negative person effect means that a person scored lower than average.

iii. The item effect shows the difficulty of the particular item. A positive item effect indicates that the item is easier than average (i.e., more people answered it correctly than the average item); a negative item effect means that the item is more difficult than average (fewer people answered it correctly than the average score across all items).

iv. Finally, the residual reflects the influence of the $p \times i$ interaction, other systematic sources of error not expressly included in the one-facet measurement, and random events.

The observed score equation for a one-facet crossed design can be written with the terms regrouped as below:

$$X_{pi} = \mu + (\mu_p - \mu) + (\mu_i - \mu) + (X_{pi} - \mu_p - \mu_i + \mu).$$

Each effect, with the exception of the grand mean has a distribution (Shavelson & Webb, 1991; Marcoulides, 2000). Shavelson and Webb (1991) and Marcoulides (2000) further explain that the grand mean is a constant and so its variance is zero. Each distribution comes with a mean of zero and variance of σ^2. The variance, σ^2, is called a 'variance component.'

Considering the person effect, the mean of the person effect over all persons is zero. This is:

$$E_p\,(\mu_{p\,-}\,\mu) = E_p\,(\mu_p) - E_p\,(\mu) = \mu - \mu = 0$$

The variance for the person effect is given in notation by σ_p^2. It is called the variance component for persons which is also termed as the universe-score variance. It is given by:

$$\sigma_p^2 = E_p\,(\mu_{p\,-}\,\mu)^2$$

This is the average of the squared deviations of the persons' universe scores from the grand mean. The variance component for persons shows how much persons differ from one another in their achievement.

In the third place, the mean of the item effect is zero. Its variance is σ_i^2. It is given by:

$$\sigma_i^2 = E_i\,(\mu_{i\,-}\,\mu)^2$$

The variance component for items shows the extent to which items differ from each other in terms of difficulty.

Finally, the fourth effect, the residual also comes with a mean of zero. Its variance is $\sigma_{pi,e}^2$. The residual variance component reflects the confounding of the $p\times i$ interaction effect with other sources of variation such as unmeasured and systematic variability which are denoted by *e*. The $p\times i$

effect reflects the fact that not all students find the same item equally difficult or easy. The *e* effect also reflects, in part, unsystematic or random error sources and systematic impacts from facets not clearly included or controlled in the one-facet G study.

The overall variance of a collection of observed scores, Xpi, over all persons and items in the universe is therefore given by the addition of the three variance components:

$$\sigma^2(X_{pi}) = \sigma^2_p + \sigma^2_i + \sigma^2_{pi,e}$$

This indicates that the variance of item scores in a one-facet crossed design can be divided into three separate sources of variation attributable to differences between persons, items and the residual. For two- and three-facet crossed designs, the variance components have been listed in the third column of Tables 2 and 3 respectively. Their explanations and deductions from them follow a similar trend as with the one-facet crossed design.

Variance components of crossed designs with fixed facets

Facets in a G study can be fixed, where the decision maker will not generalise beyond the conditions of the facets used in the G study. Statistically, G theory handles a fixed facet by averaging over the conditions of the facet (Shavelson & Webb, 1991; Brennan, 2010). For example, in the UEW- SIP, where the universe of admissible observations is defined by the two facets—raters (r) and occasions (o), if the design of the measurement procedure is assumed crossed, then it is given in notation by $p\times r \times o$. If occasion is considered as a random facet while rater is considered fixed, then the $p\times r \times o$ design becomes a mixed design (mixed model) since it contains both random and fixed facets.

In averaging over the conditions of the fixed facets, Shavelson and Webb (1991) give the following three steps.

i. Step 1: Run an analysis of variance handling all sources of variance as random to allow estimation of variance components from the fully random analysis. For example, with the $p\times r \times o$ design, even when the rater (r) facet is considered as fixed and the others random, all facets must be analysed as random first. The variance components that should be estimated in the fully random analysis are $\sigma^2_{p,}$ σ^2_{o}, $\sigma^2_{r,}$ σ^2_{po}, $\sigma^2_{pr,}$ σ^2_{ro} and $\sigma^2_{por,e.}$

ii. Step 2: Identify the random part of the mixed design and their related variance components to be computed. In the $p\times r \times o$ design with persons (p) and occasions (o) as random and raters (r) fixed, the random part of the design is persons (p) crossed with occasions (o). The variance components to be computed for the random part of the design, then, correspond to persons (p), occasions (o) and the interaction between persons and occasions, and the remaining error (*po,e*). Let these variance components be σ^2_{p*}, σ^2_{o*}, and $\sigma^2_{po,e*}$, to differentiate them from the variance components computed in the fully random design analysis in Step 1.

iii. Step 3: Calculate the variance components for the random part of the mixed design identified in Step 2. Each variance component is the variance for that source from the fully random design in Step 1 added to $1/n_{\mathrm{r}}$ multiplied by the variance component corresponding to the interaction between that source of variance and the fixed facet. The notation, n_r, gives the number of conditions of the fixed facet (r). Hence, the variance components to be estimated are as follows.

$$\sigma^2_{p*} = \sigma^2_p + \frac{1}{n_r}\sigma^2_{pr}$$

$$\sigma^2_{o*} = \sigma^2_o + \frac{1}{n_r}\sigma^2_{or}$$

$$\sigma^2_{po,e*} = \sigma^2_{po} + \frac{1}{n_r}\sigma^2_{por,e}$$

It must be noted that the right-hand side of each equation includes only interactions that include that fixed facet.

It must be noted that if a decision maker decides that it does not make conceptual sense to average over the conditions of the fixed facet, or the estimated variance components of the fixed facets ($\sigma^2_{r,}$ $\sigma^2_{pr,}$ $\sigma^2_{ro,}$ $\sigma^2_{por,e}$) as delineated in Step 1, are large, it is recommended that every condition of the fixed facet be analysed and reported separately (Shavelson & Webb, 1991). For example, in the UEW-SIP, the rater facet is made up of both school-based mentors and university supervisors. These different raters rate mentees on different number of times and occasions and also contribute differently to the overall score for grading. It therefore makes sense conceptually to analyse the two conditions of the rater facet separately.

Variance components of nested designs with random facets

Suppose a large set of mathematics test items is to be administered and the items have been divided and put into test forms such that each set of items appears on only one of the forms and different persons are administered different items. This is the situation of a one-facet nested measurement where items *i*, are nested in test forms *f* (i.e., i:f or i(f)). For an i:f design, the observed score for a person on one item (X_{pi}) can be divided into three as follows (Shavelson & Webb, 1991; Brennan, 2010).

$$X_{pi} = \mu \qquad \text{(grand mean)}$$
$$+ (\mu_p - \mu) \qquad \text{(person effect)}$$
$$+ (X_{pi} - \mu_p) \qquad \text{(residual effect)}$$

It must be noted that this design has no separate term for the item effect. The item effect is part of the residual term. This is because different persons are administered different items and thus the item effect cannot be estimated independently of the person-by-item interaction. This makes the μ_i and μ_{pi} confounded (Shavelson & Webb, 1991). The full expression of the residual effect is given by:

$$(X_{pi} - \mu_p) = (\mu_p - \mu) + (X_{pi} - \mu_i - \mu_p + \mu).$$

This expression according to Brennan (as cited in Shavelson & Webb, 1991) shows that, the item effect is part of the residual.

Just as in the crossed design, the person effect and the residual effect have distributions with mean zero and accompanied variances. The variance component of the object of measurement (persons) is the universe score variance and is given in the same way as in the crossed designs. That is:

$$\sigma_p^2 = E_p(\mu_p - \mu)^2$$

The variance component for the residual is expressed as ($\sigma_{i,pi,e}^2$) and as explained earlier, affirms the fact that the item effect is confounded with the effect for the interaction between persons and items which is confounded with unsystematic sources of variation. It is given by:

$$\sigma_{i,pi,e}^2 = E_p E_i (X_{pi} - \mu_p)^2$$

The variance for the collection of observed scores, $X_{pi,}$ for all persons and items is therefore, the addition of the two variance components (Shavelson & Webb, 1991):

$$\sigma^2(X_{pi}) = \sigma^2_p + \sigma^2_{i,pi,e}$$

In the assertion of Cronbach et al. (as cited in Shavelson & Webb, 1991), the one-facet nested design shows the disadvantage of using a nested design as compared to a crossed design in a G study. The main disadvantage is that it becomes impossible to estimate a separate variance component for the item effect. It is therefore advised that in order to estimate the largest number of sources of variability in measurement as possible, a fully crossed design must be used whenever possible.

Variance components of nested designs with fixed facets

In nested designs, the facet(s) can be fixed. Let us consider a study of 'personality construct self-concept' by Shavelson, Hubner and Stanton (as cited in Shavelson and Webb, 1991). Using the Marsh's Self-Description Questionnaire (SDQ) which assesses numerous self-concept dimensions in academic and non-academic areas, three dimensions which are general self-concept, academic self-concept and mathematics self-concept were assessed in the study. The general, academic and mathematics scales of the SDQ contain 16, 12, and 10 items each, respectively, and were administered to 140 seventh grade students. In the design of the study, items (*i*) are nested within scale(s) because there are multiple and different items on each scale. Both items and scales are crossed with persons (*p*) because each person responded to all items on all scales. The design of the study is given by p × (i:s) and it is a partially nested design because part of the design is crossed.

The three scales used in the study were chosen to assess three different levels of self-concept in the 140 pupils and no generalisations to other areas of self-concept were intended. The scale facet is therefore considered as fixed. According to Shavelson and Webb (1991), both averaging over the three scales and analysing each scale separately makes conceptual sense. This is because the dimension of self-concept measured by each scale is distinct and can be analysed separately. There could also be components of a single undifferentiated self-concept variable which is well represented by the average of the scales and so the decision maker can average over the scales.

In averaging over the three scales, three steps are followed just as in the case of crossed designs with fixed facets.

i. Step 1: Run an analysis of variance handling all sources of variance as random to allow estimation of variance components from the fully random analysis. For example, with the p × (i:s) design, even when the scale (s) facet is considered as fixed and the others random, all facets must first be analysed as random. The variance components that should be estimated in the fully random analysis are σ^2_p, σ^2_s, σ^2_i, $\sigma^2_{i:s}$, σ^2_{ps} and $\sigma^2_{pi:s,e.}$ It must be noted that to be able to create a balanced design to make the computation of variance components easier using statistical software, all the scales must have the same number of items. In this example, 10 items were sampled from each scale for the analysis.

ii. Step 2: Identify the random part of the partially nested design and their related variance components to be computed. The random part of the design has persons crossed with items—$p\times\ i$. The variance components to be computed for the random part of the design, then, correspond to persons (p), items (i) and the interaction between persons and items, and the remaining error (pi,e). The variance components are σ^2_{p*}, σ^2_{i*}, and $\sigma^2_{pi,e*}$. The asterisks differentiate these variance components from the variance components computed in the fully random design analysis in Step 1.

iii. Step 3: Calculate the variance components for the $p \times i$ design, which is the random part of the mixed design identified in Step 2. The notation, n_s, gives the number of conditions of the fixed facet (s), which is three. The variance components to be estimated are as follows.

$$\sigma^2_{p*} = \sigma^2_{p} + \frac{1}{n_s}\sigma^2_{ps}$$

$$\sigma^2_{i*} = \sigma^2_{i:s}$$

$$\sigma^2_{pi,e*} = \sigma^2_{pi:s,e}$$

Estimation of variance components

The variance of an effect in a G study is called a variance component (Shavelson & Webb, 1991). For estimation of variance components in G theory, it is recommended to use a statistical software package. There are a few of such statistical software packages available. GENOVA (Crick & Brennan, 1983) was put into the public domain in 1983 for use. It is a software suite specifically developed for generalizability analyses. It is accessible to be downloaded for free at (http://www.education.uiowa.edu/casma/computer_programs.htm#genova) for

Macintosh and PC Windows applications and has an additional handbook which provides coaching on basic and advanced applications (Lakes & Hoyt, 2009). The GENOVA computer programme provides output for variance estimates and their computed standard errors of measurement (SEM) but has the disadvantage of requiring complete data.

Another statistical software package is EduG (EDUCAN Inc., & Institute for Research and Documentation in Pedagogy [IRDP], 2010). It is specially designed for G and D studies. The EduG Computer programme provides output for variance estimates with estimated standard errors of measurement (SEM) and also partitions the error variance from each facet into relative and absolute error variances. EduG has the disadvantage of requiring complete data. The software is compressed for downloading and can be obtained freely from the IRDP website at this address: http://www.irdp.ch/edumetrie/englishprogram.htm.

Alternatives are available in the Statistical Analysis System (SAS) and the Statistical Package for the Social Sciences (SPSS) software packages that use the VARCOMP procedure (Putka & McCloy, 2008). These techniques use the normal SAS and SPSS interface. The advantage with SAS and SPSS is that they handle missing data without adverse effects on the output. The main disadvantage of these software is that they give variance component estimates without their accompanied SEM. This means that indices about the exactitude of the variance components are unavailable to be used as added information to justify their authenticity.

Universe scores

If a number of samples of measurement are taken, then for any person, the mean score for every instance can be obtained in the universe of generalisation. For such a person, the expected value of such mean scores is defined as the person's universe score. The variance of universe scores over all persons in the population is called universe score variance. It is the equivalent of the true score variance in CTT (Brennan, 1992, 2010).

Error variances

Given the various sources of variability as in Tables 1, 2 and 3, the variance components other than σ_p^2, which is the variance component for the universe score for persons, constitute one or two different kinds of error variances. These are the variations among persons in a group which are attributable to chance factors. The application of the appropriate G theory analysis would estimate the variance components and then assign them for the estimation of two main kinds of error variances which are "absolute error variance" and "relative error variance."

Absolute error variance, σ^2 (Δ)

Shavelson and Webb (1991) and Brennan (2010) assert that, when measurement is used to index an individual's or group's absolute level of performance on a measured attribute, it contributes to "absolute decisions." The variance of errors associated with these decisions are called absolute error variance.

According to Brennan (2010), absolute error is the error involved in using a testee's observed mean score as an estimate of his or her universe score. It is basically the difference between a person's observed and universe scores. Mathematically:

$$\Delta_p = X_{pRO} - \mu_p \quad (1)$$

It must be noted that the equation of the linear model for an observable mean score over n'_r raters and n'_o occasions for a D study $p \times R \times O$ design is given by:

$$X_{pRO} = \mu + \nu_p + \nu_R + \nu_O + \nu_{pR} + \nu_{pO} + \nu_{RO} + \nu_{pRO} \quad (2)$$

The terms in equation 2, with the exception of μ, are known as score effects and their variances are called D study variance components. On the presumption that the population and all facets in the universe of generalisation are infinite, these variance components are called random effects variance components.

Also, the equation of the universe score when the raters and occasions facets are random is given by:

$$\mu_p = E_R E_O X_{pRO} = \mu + \nu_{p,} \quad (3)$$

where E stands for expected value. The variance of universe scores, $\sigma^2(p)$, is denoted generically by $\sigma^2(\tau)$. From equations 2 and 3, equation (1) becomes:

$$\Delta_p = \nu_R + \nu_O + \nu_{pR} + \nu_{pO} + \nu_{RO} + \nu_{pRO} \quad (4)$$

Hence, the variance of the absolute errors, $\sigma^2(\Delta)$, is the addition of all the variance components except the universe score variance, $\sigma^2(p)$.

Hence, $\sigma^2(\Delta) = \sigma^2(R) + \sigma^2(O) + \sigma^2({}_pR) + \sigma^2({}_pO) + \sigma^2(RO) + \sigma^2({}_pRO)$

Relative error variance, σ^2 (δ)

According to Shavelson and Webb (1991), when measurements are used to rank order individuals or groups based on performance on a measured attribute it contributes to "relative decisions." The variance of errors associated with these types of decisions are called relative error variance.

Relative error is defined as the error associated with using a testeee's observable deviation score as an estimate of his or her universe deviation score. It is given by the difference between a person's observed deviation score and his universe deviation score (Brennan, 1992, 2010). Mathematically:

$$\delta_p = (X_{pRO} - \mu_{RO}) - (\mu_p - \mu), \qquad (5)$$

where μ_{RO} is the expected score over persons of the observed scores, $X_{pRO.}$ For the $p \times R \times O$ design and an infinite universe of generalisation, it can be shown that:

$$\delta_p = \nu_{pR} + \nu_{pO} + \nu_{pRO} \qquad (6)$$

The variance of these relative errors is the addition of the variance components for the three effects in equation 6. It is given by:

$$\sigma^2(\delta) = \sigma^2(pR) + \sigma^2(pO) + \sigma^2(pRO) \qquad (7)$$

Relative error variance is the analogue of an error variance in CTT.

Coefficients and indices

A reliability coefficient gives a summary of the results of a G study. For any measurement design, a generic reliability coefficient called generalizability coefficient (G coefficient) can be computed. It is given by the ratio of estimated true score variance to estimated total observed score variance. The G coefficient has a range from 0.0 to 1.0 and higher values reflect more reliable measurement procedures and vice versa (Marcoulides, 2000; Cardinet, Johnson & Pini, 2010). Shavelson and Webb (1991) and

Cardinet et al. (2010) caution that the definition of the coefficient in G theory depends on how a measurement is to be used. This is because error variance is not the same for absolute and relative decisions. Consequently, the magnitude of the G coefficient will depend on the kind of decision to be made.

Three kinds of reliability-related coefficients are available in G theory. These are a coefficient of relative measurement, a coefficient of absolute measurement, and a coefficient of criterion-referenced measurement (Cardinet et al., 2010). The coefficient of relative measurement addresses the proportion of total score variance that is attributable to the true variation among randomly sampled objects of study. It denotes the percentage of variability in individuals' obtained scores that is systematic. It gives the extent to which the measurement procedure used is able to differentiate reliably among the objects of measurement concerned.

According to Cardinet et al. (2010), this is the coefficient that Cronbach et al. (1972) defined particularly as the G coefficient. It is symbolised by $E\rho^2$ and is interpreted in the same way as the coefficient of reliability in CTT. The G coefficient, $E\rho^2$, is applicable if scores are to be given relative interpretations as in norm-referenced instances. It helps one to estimate how precisely the measurement procedure can locate the objects of measurement, say, persons, relative to one another, and to estimate correctly the intervals between them.

Mathematically, in a one-facet crossed $p \times i$, design for a G study, the $E\rho^2$ is given by the ratio of universe score variance to the addition of universe score variance and relative error variance (Shavelson & Webb, 1991; Brennan, 2010). That is:

$$E\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2\,(\delta)} \text{ or } E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{Rel}^2}$$

The coefficient of absolute measurement, defined as the dependability coefficient (D coefficient) (Brennan and Kane, 1977; Brennan, 2010), and symbolised by Φ (Phi), on the other hand, evaluates the ability of a measurement procedure to locate the objects of measurement, say, persons reliably on a scale in absolute terms (Cardinet et al., 2010). D coefficient is used in absolute interpretations, especially essential when making domain referenced decisions, and it makes use of all variance components except the object of measurement.

Mathematically, the D coefficient is given by the ratio of universe score variance to the addition of the universe score variance and absolute error variance (Brennan, 2010; Shavelson & Webb, 1991). That is:

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2\,(\Delta)} \text{ or } \Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{Abs}^2}$$

The dependability index, Φ, differs from the generalizability coefficient, $E\rho^2$, because the former includes absolute error variance, $\sigma^2(\Delta)$, while the latter includes relative error variance $\sigma^2(\delta)$. With absolute decisions, the main effect of items (attribute/construct measured)—how difficult an item is, influences absolute individual performance and so plays a key role in the definition of measurement error. Since $\sigma^2(\Delta)$ is usually larger than $\sigma^2(\delta)$, the consequential effect is that Φ is smaller in value than $E\rho^2$ (Marcoulides, 2000; Shavelson & Webb, 2005; Webb et al., 2006; Brennan, 1992, 2010).

Phi(lambda) coefficient, Φ(λ), is a coefficient of criterion-referenced measurement and extends the Phi coefficient to cover cut-off score applications Brennan (as cited in Cardinet et al., 2010). Cardinet et al. (2010) assert that, the Φ(λ) indicates how reliably a measurement instrument can locate persons' results in terms of a performance standard (cut-off score) or criterion which is set at λ on the measurement scale. For example, in the UEW-SIP, the pass mark (cut-off score) is set at 50.0 points on a 0 – 100 scale and so Φ(50) will indicate how reliably the measurement procedure of the SIP could place individual interns on one or other side of this point. In effect, it estimates the distance between the individual scores and the chosen cut-off score. It is basically, the dependability of the measured difference between an achieved score x and the cut-off score S. According to EDUCAN Inc. and IRDP (2010), the basic expression for this is: $(x - S) = (x - m) + (m - S)$,

where $(x - S)$ is the distance from the score to the threshold S,

$(x - m)$ is the difference between an individual's observed score, x and the average score of the sample of other candidates, and

$(m - S) = B$ is the difference between the average score, m, for the sample and the cut-off score, S.

According to EDUCAN Inc. and IRDP (2010), when the cut-off score for the assessment is not different from the mean of the sample values, the difference, B, $(m - S)$, gives a null value, and the reliability of the difference is given by Phi coefficient which is similar to Coef_G absolute, except for Whimbey's correction.

EDUCAN Inc. and IRDP (2010) assert that Whimbey's correction is a weighting of estimated variance components by a correction factor which takes the type of sampling of each facet into consideration in order to compute a most appropriate G coefficient.

According to EDUCAN Inc. and IRDP (2010), if $(m - S) = B$ does not give a null value, then, it must be taken as a source of true variance. Livingstone (as cited in EDUCAN Inc. & IRDP, 2010) proposes that in this case, B^2 must be added to both the numerator and denominator of the classical reliability coefficient. Brennan and Kane (1977) consent to this approach but insist that the average of the sample, m, is subject to sampling fluctuation. For this reason, they introduced a new coefficient called Phi(lambda), which subtracts this source of error variance, w, from B^2. The formula for Phi(lambda) therefore becomes (Brennan & Kane (1977):

$$\Phi(\lambda) = \frac{\sigma^2 p + (B^2 - w)}{(\sigma^2 p + \sigma^2 (Abs) + (B^2 - w)}$$

According to EDUCAN Inc. and IRDP (2010), w frequently has a higher value than B^2 and when this happens, Phi(lambda) will consequently have a lower value than Phi coefficient. They continue that, this scenario is intuitively non-valid and Phi(lambda) can even have a negative value which is at variance with the very definition of the concept of reliability. Whenever the computed Phi(lambda) is less than the Phi coefficient, EduG provides a restricted Phi(lambda) by replacing Phi(lambda) by the value of the Phi coefficient (Coef_G absolute) which is then used for interpretation.

The formulae used to calculate G coefficients remain controversial even up to today (EDUCAN Inc. & IRDP, 2010). The ρ^2 coefficient formula was used by earlier psychometricians because they made the assumption that the objects of measurement are drawn at random from an infinite population, thereby assuming a random effect model. EDUCAN Inc. and IRDP (2010) continue that when the facets (object of measurement) whose levels are to be differentiated are fixed facets such as gender, people's religious affiliations, personality types and so on, the importance of the sampling effect must be quantified by an ϖ^2coefficient and not a ρ^2.

The EduG programme resolves this conceptual problem by ensuring that the computation of the estimates of the variance of the effects concerning the object of measurement (true score variance) and the total variance (total expected observed score variance), which are used in computing the G coefficients, take the type of sampling into consideration. This may be purely random, fixed or random finite. EduG thus, computes a ρ^2, an ϖ^2 or an intermediate value for its G coefficients (EDUCAN Inc. & IRDP, 2010).

The feat described above is attained by the application of 'Whimbey's correction' to the classically obtained variance component estimates. The Whimbey's correction is given by the expression, $\frac{\mathrm{N(f)}-1}{\mathrm{N(f)}}$, where $\mathrm{N(f)}$ is the size of the facet F universe in the given design. Each ANOVA estimated variance component is then weighted by this coefficient before it is used for further computations (EDUCAN Inc. & IRDP, 2010). These corrected components are shown in columns on all EduG output of G study analysis.

Theoretical Framework

Classical test theory (CTT)

According to Traub (1997), classical test theory (CTT) is a product of work of previous years that yielded fruit in the early 20th century. The theory was based on three significant feats of the previous 150 years. The first was the acknowledgement of the presence of errors in all psychological measurements, the second was the modelling of this error as a random variable, and the third was the development of the concept of correlation and how to quantify it.

In 1904, Spearman put out logical and mathematical arguments that test scores are 'fallible measures' of human attributes and that the obtained correlation between such fallible test scores is lower than the correlation between 'true objective values.' In his bid to explain the terms 'fallible measures' and 'true objective values,' Spearman (as cited in Crocker & Algina, 1986; Traub, 1997 & Burns, 1998) developed the notion of correction of correlation coefficient for attenuation due to unreliability of measurement instruments. He also showed the method to obtain the index of reliability needed in making such correction. Crocker and Algina (1986), Traub (1997) and Burns (1998) assert that Spearman's work in principle, marked the birth of the classical true score model.

The framework of CTT was further elaborated and refined by Spearman himself, Guilford, Guttman, Gulliksen, Lord and Norvick (as cited by Crocker & Algina, 1986; Traub, 1997), and others for the half century or more following 1904. The publication of the Kuder-Richardson formulas (KR 20 & KR 21) (as cited in Crocker & Algina, 1986; Traub, 1997) laid a major landmark in 1937. The event was followed, by the introduction of the notion of lower bounds to reliability and the framework for enhanced understanding by Guttman (as cited in Traub, 1997). The developmental peak of CTT was reached in the methodical treatment it received from Novick, and Lord and Novick (as cited in Traub, 1997).

The basics of the model put forward by Spearman on CTT was that any observed score could be seen as the composite of two hypothetical components (a true score component and a random error component) (Crocker & Algina, 1986). This is mathematically given as:

$$X = T + E$$

where X represents the observed test score,

T represents the individual's true score, and

E represents a random error component.

CTT thus, partitions the observed score variance, σ_X^2, into only two components which are the true score variance, σ_T^2, and random error score variance, σ_E^2.

By the methods of estimation of reliability in CTT, it is possible to examine only a single source of measurement error at any given time. This also means that in CTT, the different sources of measurement error in a given measurement are all bulked up and undifferentiated in the single expression of the error score variance, σ_E^2.

In CTT, the concept of reliability is defined basically as the ratio of the true score variance to the observed score variance. According to Gugiu, Gugiu and Baldus (2012), an assortments of reliability estimators have been developed based on this basic definition. These conceptualisations of reliability can be categorised into three main groups. These are: stability, internal consistency, and inter-rater reliability. The first two directly represent classical definitions of reliability while the last one represents a modern measure of reliability.

According to Allen and Yen (1979), Crocker and Algina (1986) and Webb et al. (2006), stability estimators which are the test-retest method, coefficient of equivalence, and alternative forms method, are aimed at measuring the ability of a test to yield consistent results when conducted at two points in time under similar conditions. A test is declared reliable if it produces similar results, as determined by a Pearson or Spearman-rank correlation. Stability estimators estimate error of measurement due to inconsistency in forms (equivalence) and time (test-retest). It must be noted that these methods come with the problem of ensuring consistent test administration conditions and a suitable time interval.

Internal consistency estimators, cited by Gugiu et al. (2012), as split-half method and its correction to double length (Spearman-Brown, 2010), Kuder-Richardson formula (Kuder & Richardson, 1937), coefficients alpha (Cronbach, 1951) and omega (McDonald, 1999), assess the extent to which test items or a set of indicators are consistent with a designated task using a single administration of the test or survey. These estimators estimate measurement error due to inconsistency in sampling the item domain. They thrive on the assumption that the items or indicators are continuous or interval, which is not always the case. To cater for this assumption, Zumbo, Gadermann and Zeisser (as cited in Gugiu et al., 2012), contend that latest advances in measurement theory have resulted in ordinal versions of alpha and omega.

Inter-rater reliability estimators assess the extent to which given ratings from different raters grading the same essay test agree with each other. In other words, they measure the consistency with which different raters rate the same essay of a group and so estimate measurement error due to inconsistency among raters.   According to Crocker and Algina (1986) and Gugiu et al. (2012), traditionally, the proportion of agreement among raters, Cohen’s Kappa for two raters (Stemler, 2007), and multiple-rater Kappa (Fleiss, 1981) have been used mainly to measure interrater reliability. It must be noted however, that these estimators (Crocker & Algina, 1986) are informatively convincing but are conceptually different from the accepted definition of reliability and should not be considered substitutes for reliability estimates in describing an observational instrument.

The three classes of reliability estimators described so far are only apt for relative decisions and interpretations. It must be noted that only one kind of measurement error can be considered in a given analysis and each kind of estimate determines the degree to which true scores differ from obtained scores. The problem, according to Rentz (1987), Shavelson and Webb (1991), Huen and Lei (2007) and Brennan (2010), nevertheless, is that CTT does not have the ability to evaluate inconsistencies in test forms, items, raters, administrators, or occasions concurrently.

Assumptions of CTT

According Allen and Yen (1979), most of the approved procedures of creating and evaluating tests are based on assumptions related to classical true-sore theory. These assumptions describe the way errors of measurement influence observed scores. Allen and Yen (1979) add that the classical true-score model "assumes certain conditions to be true; if these assumptions are reasonable, then the conclusions derived from the model are reasonable. However, if the assumptions are not reasonable, then the use of the model leads to faulty conclusions" (p. 56 – 57).

Allen and Yen (1979) and Gulliksen (1987) give seven basic assumptions on which CTT thrives as follows. The observed (obtained), true and error scores are symbolised by X, T and E respectively.

i. Assumption 1: The observed score, X, of an examinee is the addition of two parts: T, the true score and E, an error of measurement, i.e., X = T + E. For every given examinee and test, T is presumed to be a fixed value and X and E can vary for that examinee on different testing occasions of the same test. E can be either positive or negative in value.

ii. Assumption 2: The expected value (population mean) of X is T, i.e., $\mathcal{E}$ (X) = T. This assumption means that T is the mean of the theoretical distribution of the X scores that is obtainable from repeated independent testing of the same person with the same test.

iii. Assumption 3: The error scores and the true scores obtained by a population of examinees on one test are uncorrelated, i.e., $\rho ET = 0$. This suggests that examinees with higher true scores on a test do not have systematically more positive or negative errors of measurement than examinees with lower true scores.

iv. Assumption 4: The error scores E_1 and E_2 of two different tests are uncorrelated, i.e., $\rho E_1E_2 = 0$, where E_1 is error for test 1 and E_2 is error for test 2. This suggests that if a person has a positive error score on test 1, he or she is not more likely to have a positive or negative error score on test 2.

v. Assumption 5: The error scores on one test (E_1) are uncorrelated with the true scores on another test (T_2), i.e., $\rho E_1T_2 = 0$, for two tests, test 1 and test 2, taken by the same population of examinees.

vi. Assumption 6: If two tests have observed scores X and $X^{/}$ that satisfy assumptions 1 through 5, and if for every population of examinees, T = $T^{/}$, and $\sigma_E^2 = \sigma_{E^/}^2$, then the tests are called parallel tests. Assumption 6 presents the definition of parallel tests. Thus, for a population of examinees taking two parallel tests, the true scores are equal, error variances are equal and these imply that parallel tests would have equal observed-score means and observed-score variances. To achieve parallelism, there must be same conditions of test administration for the tests.

vii. Assumption 7: If two tests have observed scores X_1 and X_2 that satisfy assumptions 1 through 5, and if, for every population of examinees, $T_1 = T_2 + C_{12}$, where C_{12} is a constant, then the tests are called essentially τ-equivalent tests. Two tests are essentially τ-equivalent if they have the same true scores except for an additive constant. For example, if two tests are essentially τ-equivalent and $C_{12} = 4$, and for the first test, three examinees have true scores 15, 16, and 17, then their true scores on the second test would be 19, 20 and 21. Parallel tests meet stronger restrictions than essentially τ-equivalent tests. For example, essentially equivalent tests may have unequal error variances and true scores may be measured more accurately by one of the τ-equivalent tests than the other. So, it means that all parallel tests meet the conditions of essentially τ-equivalent tests but not vice versa.

Comparison of classical test theory and generalizability theory

In the first place, the CTT identifies that the circumstances of a given testing situation can undoubtedly contribute to measurement error (Etsey, 2015). CTT however, evaluates these sources of error each in a separate analysis. For example, test-retest reliability assesses variations in scores across occasions, internal consistency estimate of reliability assesses the internal consistency (errors among items) of items on an assessment instrument, while the inter-rater reliability estimate assesses the consistency with which different raters rate the responses to test items by the same group of students by the use of Cohen's kappa. Unfortunately, as affirmed by Brennan (2001), Schmidt, Le and Ilies (as cited in Lakes & Hoyt, 2009), different estimates of reliability coefficients are not equivalent to one another, because each coefficient connotes a different definition of measurement error.

CTT treats variation in individual observed test scores attributable to a combination of systematic and random sources that include omitted variables, interactions between the elements of measurement and the persons measured, and brief contributions to individual performance differences that were beyond measurement interest as one undifferentiated mass called error (Shavelson & Webb, 1991). For example, for a given measure with an internal consistency reliability, say, coefficient alpha, r_{xx} = 0.70, it is explained that 70% of the variance in the observed scores indicates true differences among persons on the measured trait and the remaining 30% indicates measurement error, error which is bulked in one and undifferentiated into sources which are unidentified. Since measurement error is undifferentiated because the sources are unidentified, the effects of error attributed the various probable sources on the consistency of an estimated score is rarely considered.

G theory on the other hand, recognises explicitly, different sources of error of measurement, which are due to for instance, items, occasions, raters, and so on. Each source can be estimated as well as the effect of their interactions. The combined effect of these sources can also be assessed (Rentz, 1987). Brennan (2010) sums it up in the statement that, “G theory liberalises classical theory by employing ANOVA methods that allow an investigator to untangle multiple sources of error that contribute to the undifferentiated *E* in classical test theory” (p. 3).

In the second place, according to Cronbach et al. (as cited in Rentz, 1987), with regards to the theoretical assumptions upon which CTT and G theory work, the former emphasises classical parallel tests and assumes that measurement conditions are strictly equivalent in content, mean and inter-correlations. The difficulty in meeting this assumption of classical parallelism is that strictly parallel tests are hard to construct. Cronbach et al. (as cited in Rentz, 1987) assert that CTT does not consider sampling of the object of measurement. G theory on the other hand, is based on a less restrictive assumption of random sampling of persons and measurement conditions. These conditions could be items, occasions and raters, which must be of the same number (quantity) sampled from the same universe of generalisation to be parallel.

In the third place, each method of reliability estimation in CTT, “provides valuable information, but provides only a slice of the bigger picture” (Burns, 1998, p. 84). This means that because each source of error is analyzed independently, it achieves a single objective. Consequently, how all the various sources of error operate concurrently and fit together to affect the overall reliability of the measurement instrument cannot be accounted for (Etsey, 2015). G theory on the other hand, offers an approach for the combination of the several possible reliability indices into an all-inclusive reliability estimate and also a way of evaluating possible interactions of the sources of error.

In the fourth place, CTT estimates reliability in terms of the relative standing of individuals in a group (norm) that has been administered an assessment only. This is the case of norm-referenced measurements. On the other hand, G theory provides reliability estimates for relative decisions and absolute decisions. Absolute decisions concern how well one performed without consideration to the performance of others. It also gives accompanied relative and absolute error variances (Shavelson & Webb, 1991; Burns, 1998; Brennan, 1992, 2010).

Finally, the reliability estimates in CTT are used to evaluate the consistency of assessment results in relation to the quality of the assessment instruments and that ends it. G theory on the other hand, with G studies, is useful in the crafting of measurement designs for subsequent studies. "A systematic study of various sources of error help in developing measurement designs that reduce total error in subsequent studies" (Etsey, 2015, p. 81). G theory enables the decision maker to determine how many test items, occasions, test administrators, raters and so on, that are needed to obtain dependable (reliable) scores and best estimates of true scores.

School Internship Programmes

An internship is a period of attachment for work experience and insight which is given by an organisation to students for a short period of time. Students on an internship programme are called interns. The word internship was originally used exclusively for medical graduates, but in modern times, it is used for a wide range of placements in businesses, non-profit organisations and government establishments (Educations.com., 2015; Loretto, 2017).

Internship programmes are usually embarked on by students and graduates looking forward to acquire important skills and experience in particular fields of endeavour. Internship programmes are either arranged for students by institutions that demand them, especially when it is internship for credits or arranged by students themselves when the purpose is solely for the acquisition of practical experience (Educations.com., 2015; Loretto, 2017).

Interns may be high school students, college and university students, or post-graduate adults. Typically, an internship programme involves an exchange of services for the sake of experience between the intern and an organisation. Every intern is given an immediate supervisor who is called a mentor. In this sense, the intern becomes a mentee who is supposed to learn closely from the mentor (Masood, 2014; Educations.com., 2015; Loretto, 2017).

Types of internships

Considering the types of internship programmes that are available currently in both academia and industry, a synopsis of the following types of internships can be given (Huhman, 2011; Educations.com., 2015; Loretto, 2017).

i. Paid internships

ii. Partially-paid internships

iii. Unpaid internships

iv. Internships for credits

v. Externships

Paid internships

Paid internships are found predominantly in the private sector or large and well-established organisations which have the financial resources to remunerate students to learn while they work (Loretto, 2017). The main purposes for paid internships are to encourage the intern, (a) to bring fresh ideas and knowledge from the classroom into the company, (b) to give off their best to contribute to the growth of the establishment, and (c) to make himself/herself available for training and scrutiny on all fronts for consideration for future full-time employment (Educations.com., 2015; Loretto, 2017). It must be noted that because this kind of internship is paid for, there are strict and higher requirements that must always be met before students are accepted for such internship programmes. These internships usually last between a period of one and four months but can be longer depending on purposes, intentions and circumstances (Educations.com., 2015). The UEW-SIP is not the paid type of internship but lasts for three to four months.

Partially-paid internships

Partially-paid internships are when student trainees are paid in the form of a stipend. Stipends are usually a fixed amount of money that is paid to students on regular basis. It is a sum of money that is paid to college students to defray part of their living expenses. In instances where students are accommodated at locations that are quite distant from their places of work and are also to take care of their basic needs on their own, they are paid an amount of money that is just a token to cover such basic needs (National Association of Colleges and Employers [NACE], 2018).

My experience with some organisations in Ghana is that interns are paid an amount of money as a form of appreciation for their services rendered at the end of the internship period. In the UEW-SIP, some schools and colleges have the practice of showing their appreciation to interns in monetary terms and gifts at the end of the internship period, although this practice cannot be classified as partially paid internship since the amounts are neither regular nor fixed.

Unpaid internships

Unpaid internships are characteristically run by non-profit making charities, schools, hospitals and some Non-Governmental Organisations (NGO's) which regularly have volunteer positions (Educations.com., 2015; Loretto, 2017). Many interns make themselves available for unpaid internships in some organisations just for the reason of gaining on-the-job professional experience or academic credit requirements needed for graduation and not for financial gains (Huhman, 2011). Internships in non-profit making organisations are characteristically not the paid type, but look impressive on a resume or curriculum vitae (Educations.com., 2015).

Internships for credits

Internships for credits are situations where universities and colleges work collaboratively with companies to offer students internships for academic credits. These internships provide students with hands-on professional experience while fulfilling their so much needed academic requirements (Huhman, 2011). Internships for credits therefore, are a great way to achieve two goals at a time, which are to acquire professional experience and to accumulate academic credits. Loretto (2017) on his part adds that internships for credit require that the on-the-job learning experience is strongly related to an academic discipline to be deemed credit worthy. According to Huhman (2011), many employers offer internship opportunities for-credit only, while others offer it as an option in exchange for the candidate's time spent on the job. To the second group of employers, that is the only way they would also benefit from the internship programme.

Interns usually have to pay for the credits, in order for the employers to be able to pay the mentors for their supervisory work. These internships are arranged through the academic offices of the university or college and may last for one to two semesters in duration. To receive credits, students may be asked to keep a journal, write an essay, develop a portfolio or make a presentation about the job experience (Huhman, 2011; Educations.com., 2015).

The UEW-SIP is categorised as internship for credit. It is undertaken during the seventh semester where students are made to work in selected partnership institutions for the whole semester. Mentors of the student teachers in the schools are made to assess the lessons of the students throughout the internship programme. External supervisors from UEW also visit the schools to assess each student's work. The SIP takes four credit hours and therefore is a requirement for certification and graduation. (UEW General Agriculture

Education Re-accreditation Document, 2013; Students' Internship Handbook, 2014).

Externships

Externships, according to Educations.com. (2015), are very much related to internships with the only difference that they are of a much shorter duration. It can last from a few days to several weeks. A common name for externship is job shadowing. Though these opportunities may be for a short period, they tend to give participants practical opportunities to experience what it is like to work in particular career fields. Externships also help in providing some needed professional connections for future interaction. It is used as programme for high school or university students to discover diverse career choices so that they can decide and plan the direction for their future careers.

General characteristics of internships

For an internship programme to be successful and meet its designated objectives both on the part of the intern and the host organisations, certain characteristics must be seen in the internship programme. The following are the characteristics as identified in related literature.

1. According to the Students' Internship Handbook (2014), Masood (2014) and University of St Thomas (2018), the internship programme should contribute to the student's personal and professional development through a series of substantive and challenging work assignments and experiences. There must be adequate planning and structuring of the breadth and depth of the internship programme prior to its beginning. There should be the provision of opportunities to

apply principles and theories learned in and outside the classroom to practical and real working life situations. Realistic goals and projects should be given and outputs should be predetermined to enable students to develop a formal portfolio that they can show future employers as evidence of their work and accomplishments that culminate into a working history.

Scott Resource Group (as cited in Gates & Paul, 2004) ranked the reasons why students choose various internship programmes to participate in them, and asserts that, first on the list was "job content." Job content was explained to mean the kind of job schedules that will be able to ensure students' personal and professional development. Second on the list was "will look impressive on my resume." This is tied closely to the class and reputation of the company or establishment which offers the internship programme. The third on the list was "relevance to my degree." The fourth reason was to "learn about the field or company." It could be seen that the main reason for choice of internship is personal development. However, an effective internship programme should also benefit the organisation involved. The UEW-SIP is mainly geared toward the personal and professional development of the student. It is an integral part of the credit requirements for graduation and it is its completion that certifies that a student has fully undertaken a professional training as a teacher.

2. An effective internship programme involves a supervision component that is mentoring and educational in nature. Interns should have access to both training and supervision because of their limited experience

(Masood, 2014). Masood continues that it is unsatisfying for interns to feel lost or left out without any feeling of a sense of belongingness in an organisation. Effective training that totally guides the interns in what is expected of them in addition to an immediate supervisor that interns know they can turn to, can immensely aid the interns to succeed in their endeavours. It is important that interns are given an 'insider treatment' by their mentors to get their optimal performance. They should be offered meaningful jobs and be included in employee activities to help develop their sense of belongingness with their assigned jobs.

According to the University of St Thomas (2018), scheduled series of supervision by a principal supervisor that include regularly arranged meetings, chances for feedback, suitable opportunities for students to ask questions, meeting periodically to review progress on supervisor's and student's learning goals, idyllically facilitates a mentoring relationship and is best for an internship programme. Good mentorship is encouraged as the best way to guide students' development. Patterson (1997) and Cunningham (as cited in Gates & Paul, 2004) conclude that supervision and mentorship are key and that cautious selection and training of supervisors and mentors are imperative.

With the UEW-SIP, each intern is assigned a supervisor (mentor) who should be in the same subject area as the intern and at least at the rank of principal superintendent in the GES. It is mandatory that the mentor stays with the intern (mentee) for the entire duration of the programme (Students' Internship Handbook, 2014).

3. In the third place, an effective internship programme has a reflection and evaluation process at the close of the internship programme. There is the provision of evaluation of overall experience, provision of closure through recognition of intern contributions, reflection on learning experiences, and provision of follow-up if necessary (University of St. Thomas, 2018). Masood (2014) adds that aside a means of evaluation at the end of the internship programme, there should be consistent constructive feedback. This is because the best way to gain the most of interns' performance is to use constructive feedback to enable them to learn from their experiences. All effective internships offer structured, systematic, and constructive feedback to the interns that help them to do a thorough appraisal so that the interns may learn and grow as employees which is the ultimate aim of the internship programme.

   The UEW-SIP is well structured in this sense due to the fact that an assigned mentor does regular supervision and evaluation throughout the period with an additional supervision and evaluation by a university lecturer. The overall evaluation of an intern's work is done when a final computation of an obtained score is done and a grade has been assigned (Students' Internship Handbook, 2014).

4. There is the provision of some form of benefits or compensation to the interns (Masood, 2014). It must not always be about money. For maximum output, employers often consider offering a modest wage

instead of not compensating the interns at all. Aside compensation, the employers can offer non-financial incentives to add perquisites to the job for the new interns. These incentives may include free parking, half-day off every other Friday, free lunch or breakfast and clothing (Masood, 2014). Employers should note that even unpaid internship can be appealing if there are tangible benefits (Gates & Paul, 2004; Masood, 2014; NACE, 2018).

According to Gates and Paul (2004), Masood (2014) and NACE (2018), other non-financial incentives include housing. Not all employers can afford to meet the housing needs of interns, but a lot of gratitude is derived if any type of assistance in the direction of housing expenses is provided. If this cannot be met, assistance must be provided in locating affordable housing in the vicinity of the work. Easier availability of inexpensive housing will make the internship opportunity more exciting to prospective students, ensuring the availability of a larger number of candidates to respond to the offer of internship in a given organisation.

Availability of affordable housing has been a major problem faced by interns in the UEW-SIP. Many SHS’s do not have enough accommodation facilities for their staff on campus or elsewhere. Whenever there are financial constraints on the part of interns to obtain accommodation on their own, they turn down the offer of internship at the particular school and seek for other places where there are available housing facilities.

Characteristics of internships for credits

The following are characteristics of internships that are specific to internships for credits.

1. In the first place, for internships for credits, the programme is completed before the student graduates from the university. In this case, the timing of the internship programme is such that students undertake it close to the final year when they have done a lot of theoretical work and before the final semester of their course programmes when accumulated credits can be tendered in for processing (Gates & Paul, 2004; Masood, 2014). The UEW-SIP is undertaken in the seventh semester and so is taken before the final semester of the entire programme (Students' Internship Handbook, 2014).
2. In the second place, a successful internship programme is planned and scheduled through consultation with a department of the university or college so as to fit into an undergraduate or graduate experience (Students' Internship Handbook, 2014). The UEW-SIP is planned by the Institute for Educational Development and Extension (IEDE) of UEW in close collaboration with partnership schools and colleges and this ensures a smooth running of all the facets of the programme.
3. Finally, an effective internship programme provides strong training or orientation sessions for all students involved in the programme. This is where they will be briefed on company culture, office procedures, and code of ethics of the company, and so on (NACE, 2018; University of St Thomas, 2018). This will help bring everybody on the same page

and starting point. There should also be orientation sessions for administrators and mentors. Effective orientations ensure that everyone starts with well-defined expectations and role definitions. According to NACE (2018) and University of St Thomas (2018), the orientation, if well done, allows the effort that is put into the session to pay off throughout the programme. Supervisors and mentors are assigned to interns right after the orientation session and training of interns for specific job duties can be started or at least scheduled for the whole internship period.

NACE (2018) insists that companies should give interns a handbook and/or website of important information. A website or handbook provides guidance to students, answering frequently asked questions (FAQs) and communicating the rules and code of practice to them at all times in a warm and friendlier way and at their own convenience.

Provision of a separate intern website serves many of the purposes of the handbook, but has the added benefit of being easy to change and updating the information posted. It can be used as a communication tool, with announcements from the college relations office or even articles of interest written by the interns themselves (NACE, 2018). With the UEW-SIP, orientation sessions are always held prior to the commencement of the actual programme and it is at this time that supervisors and mentors are assigned (Students' Internship Handbook, 2014).

Benefits of internships

There are numerous benefits derived from internship programmes. These are grouped into benefits to the host organisation in general, UEW, schools and colleges involved in the UEW-SIP, and interns (students).

Benefits to the host organisation

The first benefit of internship to the host organisation is to recruit permanent workers from the interns in the nearest future. Gates and Paul (2004) stress strongly that for the host organisation to benefit from the internship programme, a close link of the internship programme to permanent hiring must always be maintained so that it will be easier getting permanent workers to recruit from the interns later for the organisation. That is, if the internship programmes are not integrated with permanent recruitment but are considered as short-term employment programmes or a way of getting work done while permanent employees are on holidays, then the possible benefits to the host organisation are much lower. Gates and Paul (2004) continue that for the host organisation to attain the best of the internship programme, the programme must attract the most desirable students. The students should also show interest in the job. In this situation, the best practice when making a selection of students for the programme is to cautiously consider organisational needs and goals so that the prospective candidate becomes the most suitable.

Patterson (1997) adding to the desirability of suitable potential candidates asserts that students recruited for intern positions should be as cautiously selected as permanent workers. Whenever host organisations choose their interns based on organisational needs and goals, they stand to gain in three ways. First, the organisation’s aim to recruit permanent employees is achieved. This is because the job schedules designated for permanent qualified employees can be done by the interns throughout the period of the internship programme. Second, employers gain from internships because they usually hire permanent employees from their best interns who have already assessed capabilities and potentials. This saves them time and money in the long run. The advantage of transforming an intern to become a full-time employee is that he/she is already familiar with the organisation, his/her position, and he/she normally needs little or no training at all. Third, because even an unpaid internship is expensive by way of training time, resources and support (Gates & Paul, 2004), it is important that the best candidate is engaged as an intern. It is totally imprudent to waste an internship on an unqualified person so long as organisational goals and recruitment requirements are concerned.

It is worthy of knowing that in the UEW-SIP, what the partner schools do in order to get the most desirable interns is that, they first consider the availability of positions for them on the staff by looking at the intern’s subject area and the possibility of getting at least eight periods to teach on the time table. Students are then taken through very short interview sessions to ascertain their capabilities. Recruitment-minded schools and colleges are able to run the internship in a recruitment-minded way so that the best interns are retained in the schools as permanent teachers after graduation.

Benefits to partnership schools and colleges

From the way the UEW-SIP is planned and executed, partnership schools and colleges stand to derive the following benefits (Students' Internship Handbook, 2014).

i. An opportunity to benefit from the special skills, new ideas, initiatives and expertise of interns.

ii. An opportunity to share experiences with interns in building a teaching portfolio, writing a philosophy of teaching statement, and conducting a school-based action research.

iii. An opportunity to identify outstanding interns for possible recruitment as permanent staff of the school.

iv. Professional training of school-based mentor teachers who will be recognised as effective and outstanding teaching professionals.

v. An opportunity for mentor-teachers and other teachers to reflect on their own teaching methods with a view to improving on their performance.

vi. Opportunity to develop a network of contacts with teachers in schools/colleges and to create learning communities.

vii. An opportunity to collaborate with UEW in the preparation of teachers.

Benefits to UEW

The UEW, which runs the internship programme in partnership with the schools and colleges, derives the following benefits (Students' Internship Handbook, 2014).

i. Enjoys a healthy and stronger collaboration with partnership schools in the preparation of its student teachers.

ii. An opportunity to improve the professional competencies of its lecturers.

iii. An opportunity to showcase the quality of its graduates as evidence of the quality of its teacher education programme.

iv. An opportunity to enlarge and enrich the context for learning for its students.

v. An opportunity to carry out joint research with teachers in the school.

It could be realised up to this point that because the UEW-SIP operates on the principles of the CSM, for the programme to be successful, all the stakeholders must play their roles as expected. The partnership schools and colleges benefit; the university also benefits; but the ultimate most desirable consequence is the preparation of a competent professional teacher for the classroom.

Benefits to students (Interns)

There are a number of ways through which students (mentees) benefit from internship programmes. The main purpose of undertaking an internship as an integral part of professional programmes is to enable students to have an opportunity to practicalise the theoretical knowledge they acquire in the classroom before graduation. When job assignments are relevant to the degree programmes being undertaken by students, it prepares the students adequately for future employment. The internships provide students with the experience that it takes to partake in a career field of interest, the professional knowhow about a particular future career, and the motivation to prepare for full-time employment after graduation. In the course of achieving this objective, professional networks and linkages are also created which can assist students later with reference letters and sureties that can help in future employment opportunities.

With internships for credit such as the UEW-SIP, it becomes the main official means by which students can have a feel of what awaits them in the world of work. With the arrangements and coordination between UEW and the partner schools and colleges, the interns are assigned only tasks that prepare them to be competent professional teachers. The Students' Internship Handbook (2014) emphasises that, it involves not just practicing teaching, but rather experiencing good practices with students in a variety of ways, with the guidance of a mentor for a full semester in a school. The SIP thus covers both teaching practice and non-teaching practices in schools to add practicality to the theory of professional education studied at the lecture theatres. Every fourth-year student who undertakes the SIP, all things being equal, must graduate having a feel of professionalism in his/her teaching career.

In a summary, it can be deduced that the UEW-SIP offers the following opportunities to interns (Students' Internship Handbook, 2014).

i. To explore and participate in a variety of aspects of school life and attain a holistic view of teaching and further insight into the complexity of educational structures and processes.

ii. To reflect upon present practice in the light of the actualities of the classroom.

iii. To learn new skills and educational practices needed to develop and maintain excellence in teaching.

iv. To be considered for recruitment by partnership schools and colleges after the internship.

v. To develop a network of contacts among teachers and schools/colleges.

vi. To cultivate the culture of classroom research as an integral ingredient of the teaching profession.

vii. To appreciate that the teaching profession involves lifelong learning.

viii. To be strengthened in their professional growth by the support from their mentors.

Limitations of Internship

First, a lot of internships do not pay at all, or pay very poorly (Mahuron, 2019). This is undoubtedly the greatest challenge for many people who consider an internship over employment. Because of various financial obligations, many adult students are not able to afford to do an internship that does not pay, or may even cost money. Mahuron continues that the United Nations, which is a major internship provider, does not pay its interns. This policy of unpaid internship is legal for non-profit organisations, but may discourage students who are economically deprived from acquiring valuable experience they otherwise need for their life careers.

In the second place, the higher requirements for some internships may go far beyond any justification to compete to work for free (Mahuron, 2019). This happens especially in organisations which are permanent-recruitment-minded at the selection phase of the internship. At times, an internship often may require a certain cumulative grade point average (CGPA) for selection and can be quite frustrating to many prospective interns.

In the third place, an internship that promises so much and raises prospective interns’ hope so high and later offers less experience becomes a major disappointment (Hyman-Parker, 1998: Mahuron, 2019). In a case like this, interns often do not get to know this until they have already committed themselves to the internship. According to Mahuron (2019), employers may not stick to the purpose of an internship. Instead of planning an informative learning experience, they may push menial tasks on the interns treating them as unskilled or temporary helping hands. Hyman-Parker (1998) concludes that this usually happens in organisations with loose organisational structures.

Finally, there is also the problem of conflicts (Hyman-Parker, 1998; Mahuron, 2019). Though internships are designed to help students with opportunities to build on their education and acquire future jobs, many conflicts may arise. Location of the job place can be a major problem. According to Mahuron (2019), moving for long distances across the country or travelling overseas for an internship can be worrying. Finding an internship locally, the organisation may not be as flexible with its programme of work as initially promised. Worse still, the internship may cause students to lose lecture times at school which would finally affect their academic performance. Hyman-Parker (1998) concludes that when there is ambivalent support from academic institutions, the above is usually the case.

Components of the UEW school internship programme

The UEW-SIP consists of school activities, teaching portfolio, statement of teaching philosophy, reflective practice and action research as its components (Students' Internship Handbook, 2014). These experiences give a holistic training to students to turn them out as competent professional teachers. They are explained below.

Under school activities, the intern is expected to participate in all phases of the professional life of a teacher. This involves:

i. classroom teaching in different contexts such as teaching in large size classes, under resourced classes, rural, urban and metropolitan areas, and mixed ability classes;

ii. observation of teaching and other activities of regular teachers of the school;

iii. observation of the teaching and other activities of other trainees and offering comments and suggestions for improvement;

iv. participation in co-curricular activities, staff meetings and other school routine assignments; and

v. interns being expected to write schemes of work, lesson plans and teaching- learning materials.

Under teaching portfolio, all interns are expected to prepare teaching portfolios to display their teaching skills, ideas, interests and other professional achievements or competencies and development. This must be done during the internship period.

With the statement of teaching philosophy, during the period of the internship, students are expected to write their philosophy of teaching statements. These are statements that reflect each intern's personal teaching values and vision. They are statements of what they believe about teaching and learning, why they hold those beliefs and how they implement those beliefs and values in their classrooms. They are also expected to state the theoretical basis or framework, extent of applicability and limitations, if any, of their teaching philosophies.

Regarding reflective practice, during the period of internship, students are expected to write their reflections on their teaching. Reflective practice is the basis of competence in teaching. This aspect of the programme stresses the importance of thoughtful analysis and continual revision of effective approaches to teaching and learning. The intern is to critically evaluate and assess how well learning objectives and outcomes have been achieved, reflect on probable reasons for non-achievement of objectives and suggest alternative approaches. The main benefit of the reflective practice for interns is that they gain a deeper insight into their own teaching styles and ultimately, greater effectiveness as teachers. Teachers who reflect are able to think creatively and critically about their teaching and to strive continuously to advance the quality of their students' learning experiences.

The last of the components of the UEW-SIP is action research. During the internship, the interns are to complete a major enquiry or a classroom action research project in their schools of practice. The project aims at helping them to experience the importance of research as an integral part of being a teacher. The projects are to address issues and questions of genuine interest and concern to the schools and communities in which they are working. The final reports of the action research are to be shared with the partnership schools or colleges and communities who have vested interest in the outcome of the study.

It could be seen clearly from the five components of the UEW-SIP that the outcome of the whole programme is to make the student teacher a competent professional teacher who is ready to brave the odds and stand out among the lots and be distinguished as such. This feat is achieved through the following goals and objectives of the UEW-SIP (Students' Internship Handbook, 2014), which are mainly to give opportunities for interns to:

i. apply and practice the principle of teaching and learning in the classroom setting and the school context;

ii. develop practical understanding and appreciation of the major teaching roles as well as the skills that are required to perform these roles;

iii. broaden their experiences, understanding and awareness of the realities of teaching and working in a school;

iv. develop an understanding of children and young people, and the skills to respond appropriately to their needs, interests and capacities;

v. develop skills in professional decision-making and capacities for reflective learning and self-evaluation; and

vi. develop professional attitudes and qualities of adaptability and sensitivity to the school and the students they teach.

The UEW and the partnership schools and colleges endeavour to ensure that the above goals and objectives become the ultimate benefit to all the interns.

Evaluation of interns in the UEW school internship programme

The evaluation and assessment of the intern are continuous processes undertaken by the mentor and the university supervisor. The evaluation is conducted throughout the period of internship. These evaluations are based on regular formal and informal observations of the intern. There are several aspects of this process. These are:

i. the evaluation of specific lessons taught by the intern. Each intern is formally rated on three occasions by the school mentor for an internal score and once by the university supervisor for an external score. The main evaluation instrument for the quality or other wise of teaching in the UEW-SIP is an observation schedule (rating scale). Three assessment scores from the mentors and one assessment score from the university lecturers are taken for each intern. The scores from the mentors are scaled down to 30% while that of the university lecturers is scaled down to 70%. The summation of these gives a total score of 100 for grading purposes and for the computation of the Cumulative Grade Point Average (CGPA) for interns. The rating scale for evaluating teaching practice is called the Intern Teaching Evaluation Form (ITEF) (see Appendix A).

ii. the evaluation of the intern's teaching portfolio, statement of teaching philosophy and reflective writings. The intern's teaching portfolio, statement of teaching philosophy and reflective writings are submitted to the university at the end of the internship period. The teaching portfolio is taken as one course and graded out of 100% while both the statement of teaching philosophy and reflective writings are taken as one course and also graded out of 100%.

iii. the evaluation of the intern's action research. The action research report is submitted at the end of the eighth semester as the undergraduate project work requirement for graduation and is graded out of 100% (Students' Internship Handbook, 2014).

The use of observational techniques in data collection

Observation is one of the oldest methods in data collection and is one of the most important techniques of social research. To Sarantakos (1993) and Gay, Mills and Airasian (2009), observation means a method of data collection that employs vision as its main technique. The focus during observation is on understanding the natural environment in which the participants live without changing it.

The idea of watching and noting events in the natural environment of the participants suggests that if the data collected through observation should be anything to rely on in a study, then they should be accurate. This demands that observers watch with the highest degree of attentiveness as possible and also record the events with the highest degree of accuracy as possible. It also implies that an observation technique becomes imperative when data must be collected in the natural settings or environment of the subjects. Observations can also be used when other techniques of data collection seem imprudent due to the nature of the participants involved in the study. For example, study of Kindergarten children may have to involve observation techniques because of the children's inability to express themselves well in an interview or inability to write well to answer items on a questionnaire. In all cases, the researcher must have a sound justification for deciding to use observation.

Types of observation

Observation can be categorised broadly into three different types. These categories are different from each other in the extent of the observer's participation in the environment (participant or non-participant); in the manner in which it is organised (structured of unstructured); and in the setting in which the observation occurs (natural or artificial) (Sarantakos, 1993; Asamoah-Gyimah & Duodu, 2007; Gay et al., 2009; McLeod, 2015).

Participant and non-participant observation

The extent of an observer's involvement in an observation ranges from no participation (non-participant observation) at all to full participation (participant observation). In a non-participant observation, observers study their subjects from outside the group without becoming part of the membership under observation or the ongoing activity. The observer is not part of the environment he/she studies, but observes and records behaviours without any interaction with the life of the setting under study. According to Gay et al. (2009), non-participant observers are less invasive and less likely to become emotionally involved with the subjects of a study than participant observers. They continued that non-participant observation becomes imperative if the observer lacks the necessary acumen to act as a true participant or if the group to be observed is too closely organised with certain well-defined characteristics that make it impossible for the observer to fit in. A typical example is that of an adult researcher who cannot be a full participant in an activity involving kindergarten children by reason of his age. The major setback of non-participant observation is that it is difficult to obtain reliable information about participants' opinions, attitudes and emotional states when the phenomena to be observed are socially sensitive, such as homosexuality and prostitution.

In participant observation, the observer becomes a part of the environment and the activity being observed, and ideally, the identity as a researcher is not known. "There are varying degrees of participant observation

— a researcher can be an active participant observer; a privileged active observer; or a passive observer" (Gay et al., 2009, p. 366).

With participant observation, the observer becomes involved in the activity while observing and collecting data on the activities, people, and the physical aspects of the setting. Opting for a participant observation may depend on the sensitive nature of the information being sought for. For example, observers who want to observe the activities of narcotic drug peddlers may themselves get involved in the drug business in order to study well the processes of the business, difficulties and attitudes from within, just as the members of the group experience themselves.

Gay et al. (2009) and McLeod (2015) point out a number of drawbacks in the use of participant observation. In the first place, the observer may lose objectivity and become emotionally involved with participants. This will undoubtedly negatively affect the consistency of data collected. In the second place, if activities to be observed move at a faster pace, some observers may encounter difficulty in participating, observing and recording at the same time. Finally, in cases where the group under study is tight-knit and closely organised, full participation may cause tension to both the researcher and the group.

Sarantakos (1993) points out that "many cases of observation lie somewhere between these two extremes of participant and non-participant observation" (p. 208). Researchers may either be more observers than participants or more participants than observers, depending on prevailing circumstances that prescribe what is expedient at given points in time.

The active participant observer is the situation where the researcher becomes an active and full participant of the phenomenon to be observed. An example is an educational researcher who is a student teacher or substitute teacher who gains access to schools and classrooms by virtue of his/her position. In this case the researcher can teach, observe his/her lessons in session and also the effect of such lessons on his students (Gay et al., 2009).

The privileged active observer on the other hand, is a situation where the researcher can move in and out of the role as a co-participant and observer. For instance, a researcher may observe accident victims at an accident scene when he is not mandatorily participating in the rescuing exercise as a Red Cross personnel. In this case the researcher can work as a volunteer Red Cross personnel and at the same time, can withdraw, stand back and watch what is happening. As a privileged observer, the researcher can move in and out of a given role and observe (Gay et al., 2009).

Finally, when a researcher assumes the role of a passive observer, he assumes no duty in the ongoing activity but rather focuses on data collection. In this case the observer is present in the observational setting, but takes no active part in the activities that go on (Gay et al., 2009).

The type of participant observation done in the supervision of the UEW-SIP is the passive observer type. This is because supervisors join the instructional delivery process in the classroom but take no part in it as it unfolds. They observe closely, record what they see in comments and scores, and discuss feedback with interns after each teaching session.

Unstructured and structured observation

With unstructured observation, an observer is placed in an environment and observes whatever he/she deems important to the research at hand. This technique is apt for the initial stage of information gathering in exploratory studies where the problem at hand has not been well-defined to enable collection of precise data. Data collected at this stage can then be used to fine-tune the problem under study. It is not strictly organised and the procedure adopted in the observation is mostly left to the observer to delineate. According to Asamoah-Gyimah and Duodu (2007), this kind of observation is more susceptible to subjective errors in both the actual observation and recordings. The observer on the spot selects certain things to observe and records, since not everything that happens at a given time can be observed.

Structured observation is used when the problem at hand has already been shaped definitely enough to enable the researcher to specify the observations to be made (Asamoah-Gyimah & Duodu, 2007). According to Sarantakos (1993), "structured observation employs a formal and strictly organised procedure, with a set of well-defined observation categories, and is subjected to high levels of control and differentiation" (p. 208). It is planned before the study begins and a printed form which can be simple or complex is used for the recording of the observation.

A major advantage of structuring observation in research is to ensure the consistency of the results even if there must be multiple observers. Worthen and Sanders (1987) add that structured and quantitative observation methods involve employing checklists or forms for recording observations which are called observation schedules. The reliability of data from structured observation lies in the sharpness of the definition of the problem under investigation and the aptness of the observers in terms of adequacy of training. The evaluation of the UEW-SIP uses structured observation. This is because what is observed is planned with an observation schedule for doing the recording.

Natural and artificial observation

The main difference between natural and artificial observation lies in the setting in which they take place (Sarantakos, 1993). The former takes place in the subjects' natural environment such as instructional delivery in a normal classroom while the latter takes place in a modified environment such as a laboratory, zoo or aquarium.

Observation in natural environment can be done either as open or hidden observation. With open observation, the participants are well informed of the nature of the study and the identity of the researcher. In hidden observation nothing is told the participants about the nature and purpose of the study and the presence of the observer. The observer can be in the natural environment of the participants and still not be noticed (Dombrowski, 2015). The UEW-SIP is a natural and open observation. The main limitation of natural observation is with the open type when the subjects are aware that they are being observed. It leads to inaccurate data as subjects may alter their behaviour. Artificial observations are mainly open observations and suffer from this same limitation.

The process of observation

Observation is time consuming and if it is used as a data collection tool, it must involve smaller samples. In using an observational technique in data collection, the first step is the selection of the topic. This is the activity to be studied by observation. It should be an observable social phenomenon. The second step involves the definition of the topic, which involves the exploration of its element and structure. This brings out the various sub-aspects of the topic that need to be dealt with (Sarantakos, 1993; Asamoah-Gyimah & Duodu, 2007).

The third step is the definition of the variables to be observed. In natural settings such as the classroom, a lot of events occur sometimes at the same time and the observer must know exactly what to observe critically and record and what to ignore. Sarantakos (1993) and Asamoah-Gyimah and Duodu (2007) explain that once the behaviour to be observed is determined and known, the researcher must clearly define what events do or do not match the intended behaviour. Once the behaviour to be observed and its sub-variables are well defined, observations must be quantified so that different observers who observe the same variable will count the same way and record the same thing. This will help ensure consistent results.

In the UEW-SIP, the major variables to be observed are categorised into four which are:

i. instructional skills, with 10 sub-variables;

ii. classroom management, with four sub-variables;

iii. communication skills, with four sub-variables; and

iv. evaluation, with four sub-variables.

The sub-variables under each of these four major variables are scored in a range of zero (0) to four (4) points (Students’ Internship Handbook, 2014).

With the variables to be observed well defined, the approach to the observation must then be determined. The approach is related to whether the study which employs observation as a data collection tool will adopt qualitative or quantitative methodology. The approach also bothers on the kind of observation, which could be structured or unstructured, participant or non-participant, and natural or artificial (Sarantakos, 1993).

Odunga (2014) on his part posits that in considering the approach to observation, ethical matters such as privacy, confidentiality or anonymity of participants and informed consent must be considered. Then after the time unit for observation is established, a decision is made on when to observe based on the purpose for which data are being collected. With the UEW-SIP, all these important indices are already pre-determined. Interns teach lessons during instructional hours and supervisors rate such lessons by observation.

The fourth step involves actual observation and recording of what is observed. All observations must be conducted systematically (Odunga, 2014). A very important decision the researcher has to take is how to classify and record the data and this must be considered at the planning stage of a given study (Sarantakos, 1993; Gay et al., 2009; Asamoah-Gyimah & Duodu, 2007; McLeod, 2015).

According to Sarantakos (1993) and McLeod (2015), classification and recording of data usually will involve three issues which are, what to record, when to record, and how to record. They assert that these issues further involve a method of sampling. The three main sampling methods are:

Event sampling: With this, the observer selects in advance what types of behaviour (events) he/she is interested in and records all occurrences of them as and when they occur. All other types of behaviour are ignored. This addresses the issue of what to observe.

Time sampling: With this, based on what to observe, the observer decides in advance that observation will take place only during specified time periods, for instance, 15 minutes every one hour, for three hours per day at certain hours in the day. He/she then records the occurrence of the specified behaviour during that period of the day only. This addresses the issue of when to observe given that what to observe has already been decided on. In natural observations, it is possible that at the designated time, the event will not occur and nothing will be recorded.

Instantaneous (target time) sampling: The observer takes a decision in advance the pre-selected moments when observation will take place and records what occurs at that instant. Any other thing that happens before or after is ignored. This method is normally used in natural and unstructured observation where the specific observational variables have not been well defined. It could be used for the refinement of the topic under study.

On the method of recording (how to record), Sarantakos (1993), Gay et al. (2009) and Odunga (2014), maintain that the mode of recording varies with the kind of observation, type of event, the intensity of information needed, and the group size. In the views of these authors, the commonest methods of recording observation are preparing field notes, writing down information in a summary using key words or codes, tape recording conversations, video recording of events and taking photographs.

Sarantakos (1993) argues that writing down information is not always feasible. Instances are where the information needed is dense, when many people are to be observed and when the identity of the observer is hidden. Also, taking notes may divert the attention of the observer to the writing and cause many events to pass unobserved. In this case, codes, which serve as symbols, shorthand recording where actions and behaviours are replaced by numerals or key words can be used and later written in detail after the observation. Tape and video recording when used make recording easier and more efficient because they produce more accurate and valid information. The only setback here is that for issues of ethics, some participants may object to their use. Still photographs can capture on-scene objects only at a given time in a motionless state and so have limited use.

An important issue that is always part of large-scale data collection using observation is the training of observers (Sarantakos, 1993; Asamoah-Gyimah & Duodu, 2007). Studies that require more than one observer and for others that the researcher is not the observer, observer training becomes unavoidable in order to determine observer agreement. The training must concentrate on issues that are central to the study and possible sources of distortion. Becker, Martin and Flick et al. (as cited in Sarantakos, 1993, p. 214) agree that concentrating on the following factors during observer training is very useful:

i. thorough understanding of the research topic;
ii. knowledge of peculiarities of the population;
iii. understanding of possible problem areas of the study;
iv. familiarities with the categories and their effective use;
v. introduction to ways of overcoming unexpected problems and conflicts;

vi. ability to follow instructions accurately and adjust them without causing bias or distortion of the data;

vii. adaptability and flexibility; and

viii. ability to observe several subjects and categories at the same time.

Dombrowski (2015), Asamoah-Gyimah and Duodu (2007) and Sarantakos (1993) emphasise the fact that training is very important even when there is a single observer to prepare him/her to become conversant with the issues enumerated above. According to Wolcot (as cited in Sarantakos, 1993, p. 214), "how to become a genuine participant observer is a difficult question, and observers only seldom reach that stage." Trainee observers should be taken through numerous pilot observational sessions after which they are made to compare their recordings. Practice sessions must be video-covered so that points of disagreement can be discussed effectively by replaying back the video severally.

In the UEW-SIP, all mentors in the partnership schools and colleges are given initial training before engagement as mentors. Continuing mentors are also trained periodically through workshops to refresh their skills. All university supervisors who are professional teachers in the first place, are also given rigorous training in supervision. Before such lecturers are assigned as internship (off-campus teaching practice) supervisors, they are first made to participate in pre-internship (on-campus teaching practice) to sharpen their skills.

Challenges in observation

According to Becker and Berger et al. (as cited in Sarantakos, 1993), Asamoah-Gyimah and Duodu (2007), Odunga (2014) and Dombrowski (2015), the major problems encountered in the use of observation in data collection are as follows.

i. The problem of observer bias.

ii. Participant reactivity.

iii. Lack of control over events and circumstances during observation.

iv. Ethical issues regarding informed consent, anonymity, confidentiality and

privacy of participants.

Observer bias refers to an observer's consistent tendency to perceive according to personal ideology and bias, producing a distorted reality. This can be contained by triangulation, where different observers are employed to overcome observer bias when the observers agree on their findings (Gay et al., 2009).

Participant reactivity may occur in participant observation as a result of participants who feel uncomfortable at the presence of the observer and for that matter alter their behaviour. To curtail this, habituation technique can be used when it is impossible for the observer to remain hidden. According to Becker (1958), habituation strategy involves exposing the subjects in a study to the participant observer for a period of time for the subjects to become used to the observer's presence. It works on the principle that the more we encounter something, the less likely we are to react to it. Here, Morrell (as cited in Odunga, 2014) stresses that it is anticipated that with the passage of time, the participants under observation will get used to the observer and start to behave naturally.

Lack of control over events and circumstances during observation happens as a result of the observational topic or variables not well defined or well understood. This results in observers becoming confused about what exactly to observe among a multiplicity of events. This is overcome when the observer chooses to observe certain events only under certain circumstances imposed by the observer. Finally, the problem of ethical issues comes up when issues on privacy, anonymity and confidentiality of participants are not addressed well. For a solution, ethical issues should always be considered and addressed as part of the design and implementation of observational method of data collection.

Empirical Review

Generalizability studies in Ghana

It is apparent that not much has been done in the area of research employing G theory in Ghana. The only work on G theory that was found in the literature is a paper by Etsey (2015). The main purpose of this paper was to explain how G theory offers an advantage over CTT in the search for true scores involving repeated measurements such as using multiple examiners to observe students' practical work or grade essays and observing examinees over repeated tasks. The general conclusion of this paper was that:

> classical true score theory has provided procedures based on reliabilities. However, these estimates have serious weaknesses due to the reliability coefficient focuses. It is therefore recommended that in the search for true scores,

> especially in multiple measurements and performance assessments, attention should be given to generalizability theory approach (p. 91).

The situation as pertains to Ghana of lack of studies in G theory is affirmed by the assertion of Jaeger (as cited in Shavelson & Webb, 1991), that:

> Generalizability methods, however, are far from pervasive in social science measurement. Indeed, articles in social science research journals contain traditional estimates of measurement reliability far more often than analyses of generalizability (assuming the researchers concerned themselves with consistency of measurement at all). One can only assume that, despite their power and promise, generalizability theory and methods are sufficiently complex that they have not yet entered the lexicon of techniques available to applied researchers in the social sciences (pp. ix-x).

Generalizability studies in some developed countries

Some available studies employing G theory in the USA and other developed countries are highlighted in the paragraphs that follow. These papers have been reviewed here in terms of their purposes, study designs, findings and conclusions.

First and foremost is a study by Froman, Owen and Daisy (as cited in Burns, 1998) which was on the usefulness of generalizability coefficients. The study was conducted in Northeastern USA and involved a sample size of 96 subjects above age 18. The purpose of the study was to assess the generalizability of people’s attitudes toward Persons With AIDS (PWA) using the AIDS Attitude Scale (AAS) for relative and absolute interpretations. The AAS consisted of two subscales, one measuring empathy, and the other avoidance behaviour. The nine-item empathy and the 12-item avoidance subscales were administered on two occasions. The design of the study was thus, a two-facet item by occasion crossed design. Acceptable generalizability coefficients were generated for both relative and absolute decisions with the two administrations of the subscales. This meant that the combined error across items and occasions did not seriously affect the assessment of attitudes toward people with AIDS.

The researchers concluded that the empathy subscale fared well under both relative and absolute decisions, across one or two observations. Also, for use on a single occasion, which is potentially more cost-effective, the avoidance subscale had its index of dependability dropping to 0.66. They therefore resolved that this subscale is less dependable if it is administered only once, especially if the scores are meant for absolute decision.

The second is a study by Burns and Froman (as cited in Burns, 1998). The aim of the study was to determine the optimum conditions (number of items and occasions) for the reliable application of a modified form of the Habitual Physical Activity Index (HPAI). The HPAI is a self-report questionnaire for assessment of physical activity which was developed by Baecke, Burema and Frijters (1982) using a sample of young adults. Burns and Froman found the HPAI to be a two-factor instrument consisting of an eight-item index for work physical activity and an eight-item index for leisure physical activity after factor analysis of responses from older American adults.

The intention of Burns and Froman was to modify the HPAI for use among older American adults.

The study adopted a fully crossed person (*p*) by item (*i*) by occasion (*o*) (i.e., $p \times i \times o$) design to allow maximum flexibility for assessing alternative designs for a D study. The object of measurement (persons), items and occasions were all assumed as random. The HPAI consisting of two 8-item indices was then administered to 45 persons on two occasions in an interval of two weeks.

The results of the study showed that for both the work and leisure indices, error was associated with the item facet and the interaction of subjects and items. The variance components for items on the work and leisure indices were 25% and 16% of the total variation, respectively. This meant that for each index, some items indicated more physical activity in the sample than did other items. Also, the variance components for the interaction of items by subjects were 27% and 37% of the total variation of the work and leisure indices, respectively. This meant that the relative standing of subjects differed from item to item. The variance components for occasions and its interactions with both subjects and items were very small and indicated that very little error was associated with occasions.

G-coefficients for both relative and absolute decisions were computed. For eight items and two occasions, G coefficients for relative decisions for the work and leisure indices were 0.86 and 0.80, respectively. G-coefficients for absolute decisions for the work and leisure indices were 0.79 and 0.75, respectively. The researchers in this study were interested in the generalizability of the HPAI when it is administered on a single occasion so as to save cost. Hence, G-coefficients were calculated for one occasion and a combination of various number of items. A G-coefficient criterion of 0.80 was set as a benchmark. For relative decisions, for the work index, a G-coefficient of 0.80 was obtained using eight items and one occasion. For relative decisions for the leisure index, 10 items were needed to bring the G-coefficient to greater than 0.80. Concerning absolute decisions, with the work index, 11 items on one occasion were needed to achieve the 0.80 criterion. For the leisure index, 13 items on one occasion were needed to bring the D-coefficient to 0.80. The developers of the modified HPAI have accordingly altered the conditions of the original HPAI to reflect the outcomes of this research.

The third is a study by Gugiu et al. (2012), who utilised G theory to investigate the reliability of grades assigned to undergraduate research papers. The primary purpose of the study was to examine the reliability of grades assigned to written project reports. The secondary purpose was to prove the use of G theory, specifically the fully-crossed two-facet model, for computing inter-rater reliability coefficients. The subjects for this study were 29 undergraduate students enrolled in an introductory-level course on Political Behaviour in Spring, 2011 at a Midwest University in the USA.

Participants were randomly assigned to one of nine groups. Two-facet fully crossed G-study and D-study designs were used whereby two raters graded four written assignments for nine student groups. This gave 72 evaluations in all. The universe of admissible observations was deemed to be random for both raters and assignments, but the universe of generalization was deemed to be mixed (i.e., random for two raters but fixed for four assignments). Four grading schemes were developed and used to evaluate the quality of each written report. Two-facet generalizability analyses were conducted to assess inter-rater reliability using a software developed by one of the authors. The main finding of the study was that there was a very high inter-rater reliability (generalizability) coefficient of 0.929 for only two raters who received no training in how to use the four grading rubrics.

The fourth is a study by Stora, Hagtvet and Heyerdahl (2013). It was based on systematic observations of family interactions that are deemed important and used for the development of the Parent Management Training Oregon (PMTO) programme which is an evidence-based parenting programme for child and adolescent conduct problems that focuses on teaching essential parenting practices to parents.

The observational data for the study were sampled from a data pool of two large studies investigating PMTO in Norway: a randomised control trial by Ogden and Hagen (as cited in Stora et al., 2013) and a study investigating the implementation process by Ogden, Forgatch, Askeland, Patterson and Bullock (as cited in Stora et al., 2013). Both studies were carried out as partnerships between the Norwegian Center for Child Behavioural Development (NCCBD) at the University of Oslo and the Oregon Social Learning Centre (OSLC). The data were collected from 2001 to 2005 from families living in Norway who sought help for child behavioural problems.

The design of the study and the analysis of data involved estimating generalizability coefficients for a measurement model in which all the raters were assumed to have rated all of the families (raters were crossed with families) and for a model in which the raters were assumed to be unique to each family (raters were nested within families). The study applied a measurement design consisting of three facets of observation: raters (r), items (i) and fathers (f). The source of variation attributed to mothers (m) served as objects of measurement in the lexicon of G-theory. Mothers were nested (:) within the facet of fathers, and so the objects of measurement was formally termed m:f. This measurement design was applied in each sub-sample of mothers for each parenting subscale. This data collection design is designated as an (m:f)×r×i design, which was interpreted that "mothers are nested within, or specific to, the two levels of the father facet." Both fathers and mothers were crossed with both items and raters, which are crossed with each other.

The aim of the study was twofold, with an overall aim being to compare the two measurement designs with regard to reliability estimates. The first aim was to assess reliability using G-theory for each of five parenting practices namely, discipline, positive involvement, problem-solving, skill encouragement, and monitoring, which adopted two measurement designs: (a) in which raters were crossed with mothers (r×m), and (b) in which raters were nested within mothers (r:m). The second aim was to estimate the number of raters needed to obtain reliable scores for each parenting practice using both measurement designs.

The findings of the study were that, "the crossed design provided higher generalizability coefficients than a nested design, implying inflated generalizability estimates if a crossed estimation model is used for a nested data collection" (p. 448). Also, three and four raters were found to be needed to obtain generalizability coefficients in the range 0.70 - 0.80 for monitoring and discipline. The study also found that one rater was sufficient for the same resultant estimate for positive involvement. Again, one rater was sufficient for an estimate in the range 0.80 - 0.90 for problem-solving. Finally, estimates of generalizability coefficients for skill encouragement were non-acceptable.

The fifth study is a research by Lakes and Hoyt (2009) on Child and Adolescent Psychology. The aims of the study were to help readers to understand the effects of measurement error on findings in clinical child and adolescent research, and also know the limitations of common methods of estimating reliability. The researchers noted that multiple sources of error are significant to most types of measures used in child and adolescent research. But, traditional reliability coefficients (estimated by CTT) which usually omit one or more relevant error sources from consideration do not provide a true indication of the effects of measurement error on study findings. They sought to achieve the aims of the study by illustrating how multiple sources of error variance, for instance, from raters and items affect the dependability of scores and to demonstrate methods for enhancing dependability of observer ratings.

For a design, the researchers made use of ratings of 181 children on child self-regulation. The design was used to illustrate the use of two-facets, which were raters and items as sources of error and three-facets, which were raters, items and occasions as sources of error. A group of trained observers (raters; n_r = 5) rated elementary school children (persons; n_p = 181) on three multi-item scales designed to measure different domains of self-regulation (n_i = 6, 7, and 3, respectively, for cognitive, affective, and physical regulation scales).

The procedure used for data collection was that students enrolled in K–5 classrooms in a private lower school were taken through a challenge course of increasing difficulty and were evaluated one at a time using the Response to Challenge Scale (RCS). This is a theory-derived, observer-rated measure of children's self-regulation in response to a physically challenging situation by Lakes and Hoyt (2009). The RCS asks raters to make inferences based on the target person's verbal and non-verbal behaviour, about his or her self-regulatory abilities in three domains, namely: physical, cognitive and affective. Five raters rated each child's responses to a challenge course using 16 bipolar adjectives, ranging from 'Distractible to Focused.' These were rated on 7-point scales. Negatively scored items were reversed before conducting generalizability analyses.

All raters received 30 minutes of training, which introduced them to indicators of strong and weak performances and equivalent ratings. The challenge course was adapted for each grade level to increase the level of difficulty for the older children. Adaptations included increasing the number and difficulty of tasks. Every child was rated by all five raters on all 16 RCS items. This made it a fully crossed design denoted as Persons × Raters × Items or p×r×i or PRI. The stability of the RCS scores was assessed at an interval of five months by rating the children on all the 16 RCS items. This introduced the occasions facet into the study. The three facets fully crossed design was denoted as Persons × Raters × Items × Occasions or p×r×i×o or PRIO.

Results of the variance components estimates from the PRI analyses in the G study for each of the three RCS subscales were used in a D study. The results showed that for a fully crossed design, for 5 raters and 6 items, the G coefficients were 0.86 for cognitive, 0.91 for affective, and 0.90 for physical. This was adjudged more dependable and economical in terms of the use of resources than the conditions of the facets that were used originally for the study which were 5 raters for 6 cognitive, 7 affective and 3 physical items which gave G coefficients of 0.86 for cognitive, 0.92 for affective, and 0.88 for physical.

For the PRIO fully crossed design, the results of the D study clearly display the importance of assessing cognitive self-regulation more especially than affective and physical self-regulation over occasions. This is shown clearly by the increases in G coefficients from 1 to 4 occasions. It was found that, using a design with 10 raters and the RCS in its current form (7, 6, and 3 items for the Cognitive, Affective, and Physical subscales, respectively), a researcher will obtain G coefficients of 0.47 for Cognitive, 0.83 for Affective, and 0.76 for Physical subscales. Increasing the number of testing occasions to four increases the expected G coefficients to 0.71 for Cognitive, 0.90 for Affective, and 0.85 for Physical subscales. Lakes and Hoyt (2009) therefore advocate that researchers who are interested in measuring behaviour traits, either they want to predict long-term outcomes or to estimate stable levels, should make use of PRIO analyses to ensure an adequate number of observations to yield reliable scores.

The sixth study is a research by Atilgan (2013) on the determination of adequate sample size required to ensure that the G and Phi coefficients obtained from a sample can be used in the estimation of the G and Phi coefficients for the population in an unbiased manner. The population for the study was 480,691 students in Turkey who took the Form A of the Social and

Behavioural Sciences (SBS) test for the 6th grade in 2008. A total of 1,200 students were randomly selected from this population and were put into 12 subgroups consisting of different sample sizes of n = 30, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1000. Each sample size was replicated 100 times in the study. The test battery contained five subtests with distinct contents and different numbers of items nested within each of the levels of the fixed facets which are Turkish, Math, Science, Social Studies and Foreign Language. All items were responded to by all the participants. The study model was defined as $p \times (i : s)$. A multivariate $p^{\bullet} \times i^{c}$ G theory design was used for the study. G and Phi coefficients were computed for the population and each of the 12 samples of different sizes. The relative root mean square error (R-RMSE) was used as the error index to evaluate the G and Phi coefficients with the G and Phi parameters estimated for the population.

The results of the study indicated that G and Phi coefficients estimated for a sample size of 30 was less than the G and Phi parameters. The R-RMSE value in this instance was greater than 0.01. When the sample size was at least 50, the R-RMSE values were less than 0.01. Thus, it was concluded that G and Phi coefficients are robust estimators of G and Phi parameters. A sample size of 400 is a more exact and robust estimator of G and Phi parameters, and increasing the sample size over 400 does not make a substantial contribution to the unbiased estimation of G and Phi parameters.

The seventh study is a research by Patrick, French and Mantzicopoulos (2020). The study aimed at evaluating the score stability of the Framework for Teaching (FFT). According to Danielson (as cited in Patrick et al., 2020), the FFT is a prominent observation instrument used most widely for teacher evaluation in the United States of America and it is made up of four domains of practice. These are preparation and planning, classroom environment, instruction, and professionalism. Two domains, which are classroom environment and instruction involve observation of teachers and hence were used for the study.

The study investigated the stability of kindergarten teachers' classroom environment, instruction, and total FFT scores for reading and mathematics lessons. Three raters each scored 200 reading and mathematics lessons taught by 20 kindergarten teachers. The design of the study was a two-facet (lessons, raters), partially nested (lessons within teachers), random design to decompose the FFT's classroom environment, instruction, and total scores into possible sources of variation (teachers, lessons, raters, and their interactions).

The findings of the study were that, for reading and mathematics, the score variances due to differences among teachers were 71% and 76% for classroom environment, 49% and 37% for instruction, and 69% and 66% for the total score respectively. On reliability estimates, G-coefficients ranged from 0.92 to 0.96 for classroom environment and total scores, and 0.87 and 0.79 for reading and mathematics instruction respectively. Decision studies concluded that two raters, each scoring three reading lessons or four mathematics lessons, are needed to achieve sufficiently reliable total scores. For scores in instruction, three raters each scoring seven reading lessons are needed and finally, more than four raters each scoring eight lessons are needed for mathematics to achieve sufficiently reliable total scores.

The eighth study is research by Ramadhan, Nasran, Utomo, Musyadad and Ishak (2019), who used G theory to design a standard instrument for assessing physics teachers' competencies. The instrument consisted of four main competencies which included pedagogical competencies, personality competencies, social competencies, and professional competencies. The research participants were 30 physics teachers in the district of Bima NTB in Indonesia and involved four experts as assessors. These assessors were experts in the field of sociology (for social competencies), research and educational evaluation (for pedagogical competencies), physics (for professional competencies), and psychology (for personality competencies).

The design of the study was a nested design for both the G study and the D-study. The G study used a two facet p x (i:r) random effects model to estimate variance components for person, rater, item, person and rater interaction, and error. The G-study design, p x (i:r), indicates the fact that items are nested within raters by grouping items under each competency under each rater. Each rater assessed different competencies, and each competency consists of four items/aspects (i).

The findings of the study were that, the major source of error was the residual (52.3%) and the variability accounted for by the object of measurement was 9.2%. The G coefficient for both relative and absolute interpretations was 0.74. The D study conducted showed that to reach a G coefficient for relative interpretation of at least 0.70 which is acceptable for research purposes (Brennan & Kane, 1977), the assessor must increase the items for each competency to four (i.e., use indicators 1, 2, 3, and 4). For a minimum G coefficient for relative interpretation ($E\rho^2$) of 0.70, the instrument must have the design, P x (I: R), where I Random, R Fixed (P = 30, R = 4 and I = 4). They concluded that, "to get the results of an assessment of authentic physics teacher competencies, the instrument developed can be used, by involving four competency indicators" (Ramadan et al., 2019, p. 336).

The ninth study is a research by Huijgen, Grift, Boxtel and Holthuis (2016), which aimed at developing a reliable observation instrument (Framework for Analysing the Teaching of Historical Contextualisation [FAT-HC]) and scoring design to evaluate the means by which history teachers promote historical contextualisation in their classrooms. Historical contextualisation is defined as the "ability to situate phenomena and individuals' actions in the context of time, historical location, long-term developments, or specific events to give meaning to these phenomena and actions" (Huijgen et al., 2018, p. 456).

The study involved five obervers (raters), five teachers (subjects of research) and 265 students in the upper track of secondary education in the Netherlands. The FAT-HC instrument comprises 48 items which evaluated four main history teaching strategies which are "reconstructing the historical context, fostering historical empathy, performing historical contextualisation to explain the past, and raising awareness of a present-oriented perspective" (Huijgen et al., 2016, p. 163). Each observer rated two video-taped lessons of each teacher making a total of 50 ratings.

The designs of the study were first, to investigate the instrument's reliability, a multivariate G study using a fully crossed (t × l × o), with history teachers (t), number of history lessons (l) and observers (o) was used. This made use of the composite of scores that ensured maximum generalizability. Second, to investigate the instrument's dimensionality, a univariate G-study at the item level which used seven facets in a crossed design was conducted.

The findings of the study were that, the teacher facet explained the largest proportion of variability (59.1%) in the observed scores which indicated a high reliability of the measurement instrument. This was followed by the residual (34.7%), the observers (4.58%), and the lessons (1.63%). Again, the item facet was responsible for most of the variance (47.25 %) in the observed scores showing that the instrument is one-dimensional with regards to evaluation of how history teachers promote historical contextualisation in their classrooms.

Finally, the researchers were concerned about the absolute level of a person's performance irrespective of others' performance and so the index of dependability coefficient (Φ) was used to identify the optimal number of observers. The benchmarks for Φ were Φ ≥ 0.70 for research purposes and Φ ≥ 0.80 for formative evaluations (Brennan & Kane, 1977). The results of a D-study showed that the ideal scoring design would use two observers who each evaluates two different lessons taught by the same teacher (Φ = 0.83) or three observers who each evaluates the same lesson taught by one teacher (Φ = 0.80).

Appraisal of Reviewed Literature

The related literature reviewed, traces the development of univariate G theory from the research findings and publication of a book on measurement theory by Cronbach et al. (1972) entitled "The dependability of behavioural measurements: Theory of generalizability for scores and profiles" (Feldt & Brennan, 1989; Burns, 1998; Brennan, 1997, 2010). According to Feldt and Brennan (1989) and Brennan (1997, 2006, 2010), the developmental process of the theoretical framework of univariate G theory were completed together with its technical report by the Cronbach team in 1960 – 1961.

Brennan (2010) posits that, the development of the theoretical framework of multivariate G theory (G theory of profiles) was started in the mid-1960s by the Cronbach team. They gave a simple but well-designed picture of the early conception of multivariate G theory. In dealing with test batteries, they emphasised the separate treatment of the scores instead of the use of composite scores.

It is worthy of noting that, it was from these early beginnings that a formidable foundation was laid for both univariate and multivariate G theory applications in research. Due to the empirically evidenced advantage of G theory over CTT in the identification and estimation of sources of measurement error (Brennan, 2010; Li et al., 2015), it is now applied in all fields of endeavour, ranging from educational assessment, psychology, special education, industry to medical practice. This is given credence by the interdisciplinary scope of coverage of the articles reviewed under the empirical review section of the current study.

The theoretical review, in addition to giving an in-depth treatment of the conceptual framework of G theory, also dealt thoroughly with CTT. The assumptions on which CTT is based, kinds of reliability coefficicients in CTT, methods of estimating reliability coefficients in CTT, and a comparison between G theory and CTT in terms of effectiveness and accuracy in identification and estimation of measurement error have been dealt with. The superiority of G theory over CTT in measurement has been clarified that G theory is an extension of CTT and also liberalises it (Rentz, 1987; Shavelson

& Webb, 1991; Burns, 1998; Brennan, 1992, 2010; Etsey, 2015). Finally, the concept of internship and the use of observational techniques in research have been well explained.

Ten empirical studies with their findings have been reviewed. The first, by Etsey (2015), was the only study found in Ghana and this paper explained how G theory offers an advantage over CTT in the search for true scores involving repeated measurements and performance assessment. It must be pointed out that, throughout the literature, empirical research involving the application of G theory in Ghana is generally non-existent. The reason is not far fetched and is given by the assertion of Jaeger (as cited in Shavelson & Webb, 1991), that, G theory methods are not pervasive in social science measurement and that G theory methods are sufficiently complex. Due to this, social science researchers are not yet exposed to the techniques involved in its application. This is an obvious gap in social science research in Ghana and attention must be paid to it.

The other nine studies reviewed outside Ghana (USA, Europe & Asia) generally border on repeated measures across occasions. They mainly focused on finding the facets that contribute most to measurement error, the effects of errors on study findings, the computation of G coefficients (mostly stability over occasions and inter-rater reliability), the index of dependability and subsequent redesigning of measurement procedures that are more reliable and economical in terms of use of resources.

A number of G theory designs have been used in these studies. The designs range from one-facet crossed random, two-facet crossed random, to combinations of crossed and nested designs, called mixed designs (two-facet partially nested design) (Etsey, 2015). This broadens the scope of design application in the studies reviewed, but it must be pointed out that, this does not exhaust the list of available designs in G theory.

Most significant of the findings of the empirical review was that of Stora et al. (2013), that compared the G coefficients of crossed and nested designs. They concluded that a crossed design provided higher generalizability coefficients than a nested design. This implied inflated generalizability estimates if a crossed estimation model is applied on a nested data collection. Another remarkable conclusion by Lakes and Hoyt (2009) is that researchers assessing cognitive self-regulation among children must measure across occasions and not just once in order to achieve stable and dependable results. Lastly, by Atilgan (2013), a sample size of 400 is a more exact and robust estimator of G and Phi parameters and that, increasing the sample size over 400 does not contribute substantially to the unbiased estimation of G and Phi parameters.

A realisation that can be said to be a weakness is that, among these nine sampled reviewed studies, only a sinlge study by Atilgan (2013) was found to have employed multivariate G theory analysis. This, points to the fact that even in the developed world, multivariate G theory methods are not yet pervasive in social science measurements. It might also be due to the fact that multivariate G theory methods are sufficiently complex. This is another obvious gap in social science research worldwide and attention must be paid to it.

CHAPTER THREE

RESEARCH METHODS

Introduction

This chapter discusses the methodology adopted in carrying out the study. The methods as described in this chapter are under six sub-sections. These are the Research Design, Population, Sample and Sampling Procedure, Research Instrument for Data Collection, Data Collection Procedure, and Data Processing and Analyses.

Research Design

A research design is defined as the detailed strategy or plan for carrying out a research study. The choice of the research design must be informed by the type of research that is being undertaken. A good choice of a design must show the basic structure of the study and objectives of the study. "The nature of the hypothesis, the variables involved and the constraints of the environment all contribute to the selection of the design" (Gay et al., 2009, p. 108).

The design used for this study was a random effect one-facet crossed design. In the UEW-SIP, interns (*p*) are made to teach on three occasions (*o*) and rated by one mentor (rater, *r*). Since the rater facet has only one level and violates a basic assumption of the application of G theory analysis, it could not be included in the G study. It was therefore treated as an unmeasured facet in the study. The design ultimately became intern (*p*) crossed with occasion (*o*), and symbolically given by ($p\times o$).

In this design, an observed score for one intern on one occasion, $X_{po,}$ can be decomposed into the following effects:

$$
\begin{array}{lll}
X_{po} = \mu & & \text{(grand mean)} \\
+ \ \mu_p - \mu & & \text{(person effect)} \\
+ \ \mu_o - \mu & & \text{(occasion effect)} \\
+ \ X_{po} - \mu_p - \mu_o + \mu & & \text{(residual)}
\end{array}
$$

The observed score equation for a one-facet crossed design can be written with the terms regrouped as below:

$$X_{po} = \mu + (\mu_p - \mu) + (\mu_o - \mu) + (X_{po} - \mu_p - \mu_o + \mu)$$

Each effect, with the exception of the grand mean has a distribution (Shavelson & Webb, 1991; Marcoulides, 2000). Each distribution has a mean of zero and a variance component σ^2. The overall variance of a collection of observed scores, *Xpo,* over all persons and occasions in the universe is therefore given by the sum of the three variance components:

$$\sigma^2(X_{po}) \quad = \quad \sigma^2_{\ p} + \sigma^2_{\ o} + \sigma^2_{\ po,e}$$

This means that the variance of occasion scores in a one-facet crossed design can be partitioned into three independent sources of variation due to differences between persons, occasions and the residual.

The design of this study is a one facet random effects model in which interns (p) were crossed with occasions (o). The standard procedures of ANOVA are used to determine the mean squares and to estimate the variance components corresponding to all sources of variation in the design. Variance components estimated for the three sources of variation are interns ($\sigma^2_{\ p}$), occasions ($\sigma^2_{\ o}$), and residual ($\sigma^2_{\ po,e}$). Table 4 shows the standard ANOVA table for the interns-by-occasions design together with corresponding computational formulae (Shavelson and Webb, 1991).

Table 4 - ANOVA Formulae for the Interns-by-Occasion Design

Source of Variation	Sums of Squares	Degrees of freedom (df)	Mean Squares	Expected Mean Squares
Interns (p)	SS_p	$n_p - 1$	$MS_p = SS_p/df_p$	$EMS_p = \sigma^2_{po,e} + n_o\sigma^2_p$
Occasions (o)	SS_o	$n_o - 1$	$MS_o = SS_o/df_o$	$EMS_o = \sigma^2_{po,e} + n_p\sigma^2_o$
Residuals(pi,e)	$SS_{pi,e}$	$(n_p - 1)(n_o - 1)$	$MS_p = SS_p/df_p$	$EMS_{po,e} = \sigma^2_{po,e}$

Source: Shavelson and Webb (1991, p. 28)

In Table 4, what is new to the usual ANOVA table is the last column for the expected mean squares. The expected mean square (EMS) is the value of the mean square that is obtainable on average by analysing repeated samples from the same population and universe using the same design. The EMS's provide weighted sums of variance components and the three EMS equations are solved to obtain estimates of each variance component. The EMS's are replaced with the corresponding observed mean squares (MS) and σ^2 replaced by $\widehat{\sigma}^2$ to give:

$$\mathrm{MS_p} = \widehat{\sigma}^2_{\mathrm{po,e}} + \mathrm{n_o}\widehat{\sigma}^2_{\mathrm{p}}$$

$$\mathrm{MS_o} = \widehat{\sigma}^2_{\mathrm{po,e}} + \mathrm{n_p}\widehat{\sigma}^2_{\mathrm{o}}$$

$$\mathrm{MS_{po,e}} = \widehat{\sigma}^2_{\mathrm{po,e}}$$

The “hat” ($\frown$), is an indicator that sample estimates are being used and not the exact estimates of σ^2. These equations are solved by using given values in the ANOVA table and working from “bottom up” to obtain the value of the residual first before the values of $\hat{\sigma}_p^2$ and $\hat{\sigma}_o^2$ (Shavelson and Webb, 1991).

The estimates of variance components obtained for occasion ($\sigma^2{}_o$), and the residual ($\sigma^2{}_{po,e}$), are then used to compute estimates for the relative and absolute error variances.

Estimated relative error variance is: $\sigma_{Rel}^2 = \dfrac{\sigma_{po,e}^2}{\acute{n}_o}$

Estimated absolute error variance is: $\sigma_{Abs}^2 = \dfrac{\sigma_o^2}{\acute{n}_o} + \dfrac{\sigma_{po,e}^2}{\acute{n}_o}$

Where $\acute{n}_o$ denotes the condition (level) of the occasion facet that is alterable in a D study.

Computed estimates for relative and absolute error variances are further used in the computation of G coefficients for relative and absolute interpretation with the formulae:

$$\text{Coef\_G relative, } E\rho^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{Rel}^2}$$

$$\text{Coef\_G absolute, } \Phi = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_{Rel}^2}$$

For a research paradigm, the positivistic paradigm (positivistism) was adopted for the study. According to Kuhn (as cited in Sarantakos, 1993), a paradigm refers to certain beliefs, values and techniques which are shared by researchers and which act as guides or maps that dictate the kinds of problems researchers should address and the types of explanations that are acceptable to them. Positivism describes reality as all things that can be perceived through the senses, independent of human consciousness, rest on order and are objective (Sarantakos, 1993). The methodology (theoretical principles of research entailed in a given paradigm) selected for the study was the quantitative methodology. This choice was driven by the kind of data (scores) to be collected, instrument for data collection (rating scale which quantifies attributes), data analysis to perform (statistical analysis) and interpretation of results. The above strategies form the research methods for the study.

Population

Amedahe (2004) asserts that, the target group about which a researcher wants to gain information and draw conclusions is known as the population. This is a group of individuals with common characteristics that are of interest to the researcher. It is called the target population. The population from which the researcher can actually select subjects for a study is termed as the accessible population (Gay et al., 2009).

In this study, the target population was all UEW regular bachelors' degree graduates from 46 departments in the 14 Faculties of UEW up to 2017/2018 academic year at all the four campuses of UEW. For the purpose of the study, the accessible population was all UEW regular bachelors' degree graduates from 46 departments in the 14 Faculties of UEW, for three consecutive academic years from 2015/2016 to 2017/2018. They were 18,339 in number (UEW 20th, 21st, & 22nd Congregation basic statistics, 2016, 2017, 2018).

The regular bachelors' degree graduates were chosen for the study because unlike their sandwich and distance counterparts, they embark on a well-planned internship programme for a whole academic term and are supervised by school mentors several occasions. I chose a three-year period for the study because I sought to study the psychometric properties of the measures over a period of time. I selected the 2015/2016 academic year as the starting point for the study because I started the study in the 2017/2018 academic year and needed the three most current academic years to work with in order to find the current state of the reliability of the measures.

Sample and Sampling Procedure

A sample is a group of individuals, items or events that is representative of the characteristics of the larger group from which it is drawn (Gay et al., 2009). The golden rule in quantitative research is to use the largest possible sample (Gall, Gall & Borg, 2007). The larger the sample the more likely the subjects' scores on the measured variables will be representative of population scores. This is what (Dunn, 2001) terms as the law of large numbers.

In this study, 9,082 bachelors' degree graduates from eight purposively selected Faculties out of 14, for the three aforementioned academic year periods were used. The eight faculties were purposively selected for the study because they are the oldest Faculties that were established before the 2015/2016 academic year and therefore had first degree graduates for all the three years chosen for the study. Because the starting point of this study was the 2015/2016 academic yaer, the eight Faculties qualified to be selected for the study. The teaching subject areas of students in these faculties are also representative of all the academic subjects that the university offers to students for teacher training. The sampled Faculties formed 57.14% of the total UEW Faculties. The sample size for the study formed 49.52% of the accessible population. Table 5 gives the distribution of students for the three academic years by the eight Faculties as at the end of 2017/2018 academic year.

Table 5 - Distribution of Students by Academic Year and Faculty

Faculty	Year			
	2015/ 2016	2016/ 2017	2017/ 2018	Total
1. Agriculture education	247	309	342	898
2. Business education	800	633	538	1971
3. Education and communication sciences	110	61	150	321
4. Foreign languages and communication	113	342	353	808
5. Social science education	577	835	1274	2686
6. Science and environment education	200	276	414	890
7. Technical education	470	490	274	1234
8. Vocational education	79	109	86	274
Total	2596	3055	3431	9,082

Source: UEW 20th, 21st, & 22nd Graduation basic statistics

The appropriateness of the sample size for the study as shown in Table 5 was based on the findings of Atilgan (2013) that when the sample size is as small as 30, the G and Phi parameters cannot be estimated stably. A sample size of 50 to 300 can be judged sufficient for the robust estimation of G and Phi coefficients. A more exact and robust estimation requires a sample size of 400, and that if the sample size is increased over 400, there is no significant input to the unbiased estimation of G and Phi coefficients. This means that the sample sizes for the Faculties and the totals for the academic years are apt for the study.

Instrument for data collection

Existing records of mentors' internship scores of students sampled for the study were used for the study. The scores were obtained from the use of an already existing observation schedule (rating scale) which is used for evaluating teaching practice. This is the ITEF shown in Appendix A. The ITEF was developed by a team of teaching methodology and evaluation experts drawn from the Faculty of Educational Studies of UEW. The guiding principles for its development were the optimum requisite skills needed to be exhibited by a teacher during instructional delivery. There is no recorded evidence yet of the psychometric properties of the scale. (Report of UEW Internship Planning Committee, 2010).

The ITEF is divided into five segments and each has a number of sub-elements that are scored on a five-point scale ranging from zero (0) to four (4). The first segment is on "Planning and Preparation" and has a maximum score of 12. This segment bothers on lesson planning and preparation with selection of appropriate teaching and learning materials (TLM's) for the lesson and comes before actual instructional delivery commences. There are three sub-elements here.

The second segment deals with "Instructional Skills" with a maximum score of 40. This is where practical instructional delivery starts in the classroom and the supervisor begins to evaluate the student's lesson as it unfolds. There are ten sub-elements here that the supervisor must listen to attentively and observe critically in order to award scores. The third segment centres on "Classroom Management" with a maximum score of 16. It bothers on the kind of relationship that exists between the teacher and the students and how he/she uses this relationship to manage his/her classroom during instructional hours. There are four sub-elements here that the supervisor must listen attentively and observe critically in order to award scores.

The fourth segment is on "Communication Skills" with a maximum score of 16. This bothers on the mode of communication between the teacher and the students through which instructional delivery occurs. It therefore demands listening with rapt attention to what the teacher says and also, observing critically what he/she writes on the chalkboard in order to assess their correctness. There are four sub-elements under this segment. The last segment is "Evaluation" with a total score of 16. There are also four sub-elements under this segment. They demand that the rater becomes attentive to both the informal and formal, formative and summative evaluation strategies of the teacher right from the beginning to the end of the lesson. It demands matching the instructional objectives to the evaluation questions in the lesson and what the teacher does both in the course of teaching and after teaching to see whether each instructional objective is fully evaluated. The total score for the rating of each lesson using the ITEF is 100.

Data Collection Procedure

To deal with ethical issues because of the sensitive nature of academic achievement scores, I applied for ethical clearance from the Institutional Review Board (IRB), University of Cape Coast, and was given clearance (see Appendix A). An introductory letter (See Appendix B) was collected from the Department of Education and Psychology, University of Cape Coast, to introduce and grant me access to UEW files. I took the introductory letter to the Institute for Teacher Education and Continuing Professional Development

(ITECPD) of UEW for permission to have access to the teaching practice results of the selected Faculties from 2015/2016 to 2017/2018 academic years. For each student, the three scores for the three occasions from the mentors were collected. Two working months were used for data collection.

Data Processing and Analysis

In the G theory analysis carried out, the thematic course area offered by each faculty was identified and used instead of the Faculty names. Each thematic course area has similar underlying theoretical concepts, principles and methods of teaching that are the main focus for evaluation on each occasion of lesson delivery. With each thematic course area having similar underlying theoretical concepts, principles and methods of teaching, students' scores for given thematic course areas could be put together and analysed meaningfully. These thematic course areas are therefore shown on the output sheets of the analysis. Table 6 gives the breakdown of the thematic course areas.

Table 6 – Faculties and Thematic Course Areas

Faculty	Thematic Course Area
Agriculture Education	Applied science (Animal Science, Crop and Soil Science, Agricultural Economics and Extension, Agricultural Mechanization)
Business Education	Business (Accounting, Management)
Education and Communication Sciences	English and Communication (English Language, Secretariat)
Foreign Languages and Communication Education	Foreign Languages (English, French, Linguistics)
Social Science Education	Social Science (Geography, Economics, History, Social Studies, Political Science)
Science and Environment Education	Natural Science (Physics, Biology, Chemistry)
Technical Education	Technical Education (Construction, Auto-Mechanic, Electrical, Woodwork, ICT)
Vocational Education	Vocational Education (Catering and Hospitality, Clothing and Textiles)

Source: Academic Faculties, UEW Registry (2019)

A univariate generalizability analysis was performed using EduG version 6.1 statistical programme (EDUCAN Inc. & IRDP, 2010). Variance components were estimated for the person and occasion facets and their interactions for each thematic course area for each Faculty for each academic year. G coefficients for both relative ($E\rho^2$) and absolute (Φ) interpretations were computed. Phi(lambda) [$\Phi(\lambda)$] as a dependability coefficient which increases the estimate of the true variance as a function of the distance from the grand mean of the observed scores to the cut-off score was also computed for each thematic course area for each Faculty for each academic year.

Research question 1 was answered by examining the values of Coef_G relative ($E\rho^2$), Coef_G absolute (Φ) and Phi(lambda) [$\Phi(\lambda)$] for the thematic course areas for each academic year to come out with the range of values for $E\rho^2$, Φ and $\Phi(\lambda)$. This enabled me to give a verbal description of the degree of stability of the results across the three occasions of rating and the degree of dependability of generalising the results from the three occasions to the average score that the interns would have received under all the possible conditions of the occasion facet.

Research question 2 was answered by examining the percentages for the estimated variance components for the object of measurement (p), occasion (o) and the interaction between the person and occasion facets (p x o) to come out with the largest proportion for each thematic course area for each academic year. In thematic course areas where the universe score variance explained most of the variability in the observed score variance, it indicated a higher reliability of the measurement instrument (Huigen, Grift, Boxtel & Holthuis, 2016). The universe score variance serves as a differentiation variance and so is not a source of error. The identified error variance components estimated are that of occasion (o) and the p x o interaction. An examination of the percentages of the estimated variance components for these effects then revealed the major sources of error in the mentors' scores.

Research question 3 was answered by performing a D study (optimization) with the results of the G study in which the conditions of the occasion facet were varied from one to six to produce new values of Coef_G relative ($E\rho^{2*}$) and Coef_G absolute (Φ^{*}). The frequency of occurrence of cases (G coefficients of at least 0.80 for given numbers of occasion) was used as a yardstick for arriving at conclusions on optimum number of occasions. A given number of occasions was judged as optimum for a given academic year if more than half (i.e., more than four) of the eight thematic course areas attained Coef_G relative ($E\rho^{2*}$) and Coef_G absolute (Φ^{*}) of at least 0.80 at that number of occasions. That is, a G-coefficient criterion of 0.80 was set as a benchmark (Burns, 1998; Cardinet et al., 2010).

According to Brennan and Kane (1977), the G coefficients should be at least 0.70 for research purposes, at least 0.80 for formative evaluations, and at least 0.90 for summative evaluations. Webb et al. (2006) put it that, "coefficients at or above 0.80 are considered sufficiently reliable to make decisions about individuals based on their observed scores, although a higher value, perhaps 0.90, is preferred if the decisions have significant consequences" (p. 1). G coefficient of at least 0.80 was taken as the criterion because the mentors assess the interns formatively with the intention of improving teaching standards. Also, to accept a given number of occasions as optimum, I set a criterion that more than four of the eight thematic course areas should attain Coef_G relative ($E\rho^{2*}$) and Coef_G absolute (Φ^{*}) of at least 0.80 because "more than half" connotes a simple majority on which a decision can be based. This decision could then be generalised to cover the other thematic course areas for a given academic year.

Research question 4 was answered by comparing the optimum number of occasions for dependable results for each academic year with the results of a D study (optimization) of the scores of all the three academic years combined. This enabled me to come out with the optimum number of occasions for dependable results for the UEW-SIP based on the data for the three academic years used for the study.

CHAPTER FOUR

RESULTS AND DISCUSSION

Introduction

The main purpose of this study was to determine the dependability of the mentors' results of the UEW-SIP, using G theory. This chapter presents the results of the analyses of data and discussion of the findings of the study. Data were analysed by performing a univariate generalizability analysis using EduG version 6.1 (EDUCAN Inc. & IRDP, 2010) statistical programme.

Results

The study was carried out in UEW, Ghana, using eight out of 14 academic faculties with 35 departments. A total of 9,082 bachelor's degree graduates' results for the academic years from 2015/2016 to 2017/2018 were used for the study.

Research Question 1

How reliable are the results of the UEW-SIP from mentors for each academic year from 2015/2016 to 2017/2018?

Research question 1 sought to find out the extent of reliability of the mentors' results of the UEW-SIP for the three consecutive academic years from 2015/2016 to 2017/2018. Appendices E_1, E_2 and E_3 give the outputs of the G study analyses for 2015/2016, 2016/2017 and 2017/2018 academic years respectively.

Table 7 shows the G coefficients (relative and absolute) as given by the EduG statistical programme, for the mentors' results of the UEW-SIP for each

Faculty (thematic course area) and academic year from 2015/2016 to 2017/2018.

Table 7 - Generalizability Coefficients for Mentors' Results of UEW-SIP from 2015/2016 to 2017/2018 Academic Years

	Academic Year					
	2015/2016		2016/2017		2017/2018	
Specialism	$E\rho^2$	Φ	$E\rho^2$	Φ	$E\rho^2$	Φ
Applied Science	*.74*	.74	.66	.66	.72	.71
Business	.77	.70	.81	.78	.73	.71
English and Communication	.77	.64	.73	.71	.76	.73
Foreign Languages	.75	.75	.81	.81	.77	.77
Natural Science	.78	.78	.82	.81	.81	.81
Social Science	.67	.67	.68	.68	.67	.67
Technical	.69	.59	.76	.72	.76	.71
Vocational	.75	.72	.84	.81	.73	.70

Source: UEW internship scores (2019)

From Table 7, it could be seen that for the three occasions of rating for the three academic years, the G coefficient, $E\rho^2$, (relative interpretation) for 2015/2016 ranges from 0.67 for Social Science to 0.78 for Natural Science. For 2016/2017, it ranges from 0.66 for Applied Science to 0.84 for Vocational

Education. For 2017/2018, it ranges from 0.67 for Social Science to 0.81 for Natural Science.

For Phi coefficient, Φ, (i.e., absolute interpretation) on the other hand, for 2015/2016, it ranges from 0.59 for Technical Education to 0.78 for Natural Science. For 2016/2017, it ranges from 0.66 for Applied Science to 0.81 Natural Science, Foreign Languages and Linguistics and Vocational Education. For 2017/2018, it ranges from 0.67 for Social Science to 0.81 for Natural Science.

The Phi(lambda) coefficients, Φ(λ), which is actually Φ(50), from the G theory analysis are shown in Table 8.

Table 8 - Phi(lambda) Coefficients for Mentors' Results of UEW-SIP from 2015/2016 to 2017/2018 Academic Years

	Academic Year		
Specialism	2015/2016 (Φ[50])	2016/2017 (Φ[50])	2017/2018 (Φ[50])
Applied Science	0.98	0.98	0.99
Business	0.99	0.99	0.99
English and Communication	0.99	0.99	0.99
Foreign Languages	0.99	0.99	0.99
Natural Science	0.99	0.99	0.99
Social Science	0.99	0.99	0.99
Technical	0.99	0.99	0.99
Vocational	0.99	0.99	0.99

Source: UEW internship scores (2019)

From Table 8, the values of Φ(λ) range from a minimum of 0.98 for Applied Science in the 2015/2016 and 2016/2017 academic years to a maximum of 0.99 for the rest of the thematic course areas in the academic years.

It could therefore be concluded that for each academic year from 2015/2016 to 2017/2018, for relative interpretation, the mentors' results of the UEW-SIP were strongly reliable as the coefficients ranged from 0.66 to 0.84. For absolute interpretation, the results were moderately (Technical Education) to strongly dependable (all other thematic course areas) for 2015/2016 while strongly dependable for 2016/2017 and 2017/2018 academic years. Hence, using Coef_G absolute (Φ), the results were moderately dependable for Technical Education and strongly dependable for all other thematic course areas for 2015/2016 academic year, while strongly dependable for all thematic course areas for 2016/2017 and 2017/2018 academic years. For absolute interpretation using Φ(λ), the results were strongly dependable (0.98 to 0.99) for all thematic course areas.

Research Question 2

What are the identified major sources of error in the mentors' results of the UEW-SIP for each academic year from 2015/2016 to 2017/2018?

Research Question 2 sought to examine the estimated variance components of all the presumed sources of error from the design of the study to find the major sources of error in the mentors' results of the UEW-SIP for each academic year from 2015/2016 to 2017/2018.

In the 24 ANOVA Tables of analyses that follow, the estimated variance component for interns (estimated universe score variance), represents true differences among the interns (p) in terms of differential levels in intelligence, knowledge of subject matter and skills of teaching and so this variance is not an error variance. The estimated variance component for occasions (o) reflects measurement error due to differential occasional conditions of delivery and assessment of teaching. The estimated variance component due to the residual reflects measurement error due to the p x o interaction and unidentified or random sources. Hence, the two error variances are that of occasions (o) and the residual.

Table 9 gives the ANOVA table of estimates of variance components for Applied Science for the 2015/2016 academic year.

Table 9 - ANOVA Estimates of Variance Components for Applied Science for 2015/2016 (Design: p x o)

Source of	Sum of	Df	Mean	Variance	% of Total
Variation	Squares		Squares	Component	Variance
Interns (p)	27046.62	246	109.95	27.19	48.8
Occasions (o)	123.93	2	61.96	0.34	0.2
Residual (po,e)	13964.07	492	28.38	28.38	51.0
Total	41134.62	740			100

Source: UEW internship scores (2019)

It could be seen from Table 9 that, the estimated variance component for interns (estimated universe score variance), 27.19, forms as much as 48.8% of the total variance. The estimated variance component for occasions accounts for only 0.34 or 0.2% of the total variance. The largest variance component is the residual which is 28.38 or 51.0% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error, therefore, in Applied Science for the 2015/2016 academic year is the p x o interaction and unidentified sources. The occasion facet contributes so smaller a percentage (0.2%) to total variance.

Table 10 gives the ANOVA table of estimates of variance components for Business Education for 2015/2016 academic year.

Table 10 - ANOVA Estimates of Variance Components for Business Education for 2015/2016 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	48858.43	799	61.15	15.64	44.2
Occasions (o)	8889.80	2	4444.90	5.34	15.6
Residual (po,e)	22756.20	1598	14.24	14.24	40.2
Total	80504.43	2399			100

Source: UEW internship scores (2019)

It could be seen from Table 10 that, the estimated variance component for interns, 15.64, accounts for the largest proportion of 44.2% of the total variance. The estimated variance component for occasions accounts for 5.34 or 15.6% of the total variance. The next variance component is the residual which is 14.24 or 42.2% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. It follows therefore, that, the major source of error in Business Education for the 2015/2016 academic year is the p x o interaction and unidentified or random sources and this is more than twice the error margin contributed by the occasion facet.

Table 11 gives the ANOVA table of estimates of variance components for English and Communication 2015/2016 academic year.

Table 11 - ANOVA estimates of variance components for English and Communication for 2015/2016 (Design: p x o)

Source of variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	7503.42	109	68.84	17.63	36.8
Occasions (o)	3187.90	2	1593.95	14.35	29.9
Residual (po,e)	3474.77	218	15.94	15.94	33.3
Total	14166.77	329			100

Source: UEW internship scores (2019)

From Table 11, the estimated universe score variance, 17.63, which is the largest, accounts for 36.8% of the total variance. The variance component for occasions accounts for 14.35 or 29.9% of the total variance. The variance component for the residual is 15.94 or 33.3% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error, therefore, in English and Communication for the 2015/2016 academic year is the p x o interaction and unidentified or random sources and is followed closely by the occasion facet.

Table 12 gives the ANOVA table of estimates of variance components for Foreign Language and Linguistics for 2015/2016 academic year.

Table 12 - ANOVA estimates of variance components for Foreign Language and Linguistics for 2015/2016 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	13945.63	112	124.51	31.02	49.5
Occasions (o)	97.71	2	48.85	0.15	0.2
Residual (po,e)	7.43.63	224	31.44	31.44	50.2
Total	21086.96	338			100

Source: UEW internship scores (2019)

From Table 12, the estimated variance component for interns, 31.02, contributes as much as 49.5% to the total variance. The estimated variance component for occasions accounts for only 0.15 or 0.2% of the total variance. The largest variance component is the residual which is 31.44 or 50.2% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error, therefore, in Foreign Language and Linguistics for the 2015/2016 academic year is the p x o interaction and unidentified or random sources while the occasion facet contributes a very small percentage (0.2%).

Table 13 gives the ANOVA table of estimates of variance components for Natural Science for 2015/2016 academic year.

Table 13 - ANOVA Estimates of Variance Components for Natural Science for 2015/2016 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	22055.84	199	110.83	28.94	54.6
Occasions (o)	74.92	2	37.46	0.07	0.1
Residual (po,e)	9552.41	398	24.00	24.00	45.3
Total	3168.17	599			100

Source: UEW internship scores (2019)

Table 13 shows that, the estimated variance component for interns, 28.94, accounts for as much as 54.6% of the total variance and it is the largest. The estimated variance component for occasions accounts for only 0.07 or 0.1% of the total variance. The second largest estimated variance component is the residual which is 24.00 or 45.3% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error, therefore, in Natural Science for the 2015/2016 academic year is the p x o interaction and unidentified or random sources. The occasion facet contributes nearly nil (0.1%) to total variance.

Table 14 gives the ANOVA table of estimates of variance components for Social Science for 2015/2016 academic year.

Table 14 - ANOVA Estimates of Variance Components for Social Science for 2015/2016 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	47873.16	576	83.11	18.61	40.0
Occasions (o)	762.52	2	381.26	0.61	1.3
Residual (po,e)	31414.82	1152	27.27	27.27	58.6
Total	80050.49	1730			100

Source: UEW internship scores (2019)

Table 14 shows that, the estimated variance component for interns, 18.61, forms as much as 40.0% of the total variance. The estimated variance component for occasions accounts for only 0.61 or 1.3% of the total variance. The estimated variance component for the residual is 27.27 or 58.6% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error, therefore, in Social Science for the 2015/2016 academic year is the p x o interaction and unidentified or random sources while the occasion facet contributes a smaller proportion of 1.3%.

Table 15 gives the ANOVA table of estimates of variance components for Technical Education for 2015/2016 academic year.

Table 15 - ANOVA Estimates of Variance Components for Technical Education for 2015/2016 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	26012.80	469	55.46	12.77	32.6
Occasions (o)	8660.53	2	4330.26	9.18	23.5
Residual (po,e)	16097.47	938	17.16	17.16	43.9
Total	50770.80	1409			100

Source: UEW internship scores (2019)

From Table 15, the estimated variance component for interns (p), 12.77, accounts for 32.6% of the total variance. The estimated variance component for occasions accounts for 9.18 or 23.5% of the total variance. The largest variance component is the residual which is 17.16 or 43.9% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. Therefore, the major source of error in Technical Education for the 2015/2016 academic year is the p x o interaction and unidentified or random sources. The occasions facet follows in terms of error of measurement, with 23.5% of total variance.

Table 16 gives the ANOVA table of estimates of variance components for Vocational Education for 2015/2016 academic year.

Table 16 - ANOVA Estimates of Variance Components for Vocational Education for 2015/2017 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	10213.29	78	130.94	32.76	46.0
Occasions (o)	983.91	2	491.95	5.81	8.2
Residual (po,e)	5096.76	156	32.67	32.67	45.9
Total	16293.96	236			100

Source**:** UEW internship scores (2019)

From Table 16, the estimated variance component for interns, 32.76, forms as much as 46.0% of the total variance. The estimated variance component for occasions accounts for only 5.81 or 8.2% of the total variance. The estimated variance component for the residual is 32.67 or 45.9% of total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error, therefore, in Vocational Education for the 2015/2016 academic year is the p x o interaction and unidentified or random sources. The occasions facet contributes nearly only one-tenth of total variance.

It could be concluded that for the mentors' results of the UEW-SIP for the 2015/2016 academic year, putting the estimated variance component for interns (p) aside, the major source of error is the p x o interaction and unidentified or random sources and is followed by the occasion facet.

Table 17 gives the ANOVA table of estimates of variance components for Applied Science for 2016/2017 academic year.

Table 17 - ANOVA Estimates of Variance Components for Applied Science for 2016/2017 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	46005.58	308	149.37	32.96	39.2
Occasions (o)	424.01	2	212.00	0.52	0.6
Residual (po,e)	31107.33	616	50.50	50.50	60.1
Total	77536.91	926			100

Source**:** UEW internship scores (2019)

It could be seen from Table 17 that, the estimated variance component for interns, 32.96, accounts for 39.2% of the total variance. The variance component for occasions accounts for only 0.52 or 0.6% of the total variance. The largest variance component is the residual which is 50.50 or 60.1% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error, therefore, in Applied Science for the 2016/2017 academic year is the p x o interaction and unidentified or random sources. The occasion facet contributes a smaller percentage (0.6%) to total variance.

Table 18 gives the ANOVA table of estimates of variance components for Business Education for 2016/2017 academic year.

Table 18 - ANOVA Estimates of Variance Components for Business Education for 2016/2017 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	48906.00	632	77.38	20.88	53.5
Occasions (o)	4337.82	2	2168.91	3.40	8.7
Residual (po,e)	18616.85	1264	14.93	14.73	37.7
Total	71860.67	1898			100

Source: UEW internship scores (2019)

Table 18 shows that, the estimated variance component for interns (p), 20.88, contributes as much as 53.5% to the total variance. The variance component for occasions accounts for 3.40 or 8.7% of the total variance. The next estimated variance component is that of the residual which is 14.73 or 37.7% of total variance. It is second largest. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error in Business Education for the 2016/2017 academic year, therefore, is the p x o interaction and unidentified or random sources. The occasion facet contributes only nearly one-tenth of variability to total variance.

Table 19 gives the ANOVA table of estimates of variance components for English and Communication for 2016/2017 academic year.

Table 19 - ANOVA Estimates of Variance Components for English and Communication 2016/2017 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	4881.77	60	81.36	19.68	44.7
Occasions (o)	286.57	2	143.28	1.98	4.5
Residual (po,e)	2680.10	120	22.33	22.33	50.8
Total	7848.44	182			100

Source: UEW internship scores (2019)

Table 19 shows that, the variance component for interns (p), 19.68, accounts for 44.7% of the total variance. The variance component for occasions accounts for only 1.98 or 4.5% of the total variance. The largest estimated variance component is the residual which is 22.33 or 50.8%. It reflects measurement error due to the p x o interaction and unidentified or random sources. This is about half of the total variance. Hence, the major source of error in English and Communication for the 2016/2017 academic year is the p x o interaction and unidentified or random sources. The occasion facet contributes only nearly one-twentieth of variability to total variance.

Table 20 gives the ANOVA table of estimates of variance components for Foreign Languages and Linguistics for 2016/2017 academic year.

Table 20 - ANOVA Estimates of Variance Components for Foreign Languages and Linguistics for 2016/2017 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	43189.19	341	126.65	34.15	58.3
Occasions (o)	190.34	2	95.17	0.21	0.4
Residual (po,e)	16516.33	682	24.22	24.22	41.3
Total	59895.86	1025			100

Source: UEW internship scores (2019)

From Table 20, the estimated variance component for interns (p), 34.15, forms as much as 58.3% of the entire variance. The estimated variance component for occasions accounts for only 0.21 or 0.4% of the total variance. The second largest estimated variance component is that of the residual which is 24.22 or 41.3% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error, therefore, in Foreign Language and Linguistics for the 2016/2017 academic year is the p x o interaction and unidentified or random sources. The occasion facet contributes a smaller percentage (0.4%) to total variance.

Table 21 gives the ANOVA table of estimates of variance components for Natural Science for 2016/2017 academic year.

Table 21 - ANOVA Estimates of Variance Components for Natural Science for 2016/2017 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	35126.25	275	127.73	34.79	58.9
Occasions (o)	549.49	2	274.74	0.91	1.5
Residual (po,e)	12845.18	550	23.35	23.35	39.5
Total	48520.91	827			100

Source: UEW internship scores (2019)

It could be seen from Table 21 that, the estimated variance component for interns (p), 34.79, contributes as much as 58.9% to the total variance, but this does not contribiute to errors of measurement. The variance component for occasions accounts for only 0.91or 1.5% of the total variance. The second largest estimated variance component is the residual which is 23.35 or 39.5% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. Hence, the major source of error, in Natural Science for the 2016/2017 academic year is the p x o interaction and unidentified or random sources. The occasion facet contributes a smaller percentage (1.5%) to total variance.

Table 22 gives the ANOVA table of estimates of variance components for Social Science for 2016/2017 academic year.

Table 22 - ANOVA Estimates of Variance Components for Social Science for 2016/2017 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	70628.00	834	84.69	19.35	41.8
Occasions (o)	501.95	2	250.97	0.27	0.6
Residual (po,e)	44420.05	1668	26.63	26.63	57.6
Total	115550.00	2504			100

Source: UEW internship scores (2019)

Table 22 shows that, the estimated variance component for interns (p), 19.35, accounts for 41.8% of the entire variance. The variance component for occasions accounts for only 0.27 or 0.6% of the total variance. The largest estimated variance component is that of the residual which is 26.63 or 57.6% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. Hence, the major source of error in Social Science for the 2016/2017 academic year is the p x o interaction and unidentified or random sources. The occasion facet contributes a smaller percentage (0.6%) to total variance.

Table 23 gives the ANOVA table of estimates of variance components for Technical Education for 2016/2017 academic year.

Table 23 - ANOVA Estimates of Variance Components for Technical Education for 2016/2017 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	48209.20	489	98.59	25.10	46.6
Occasions (o)	5395.41	2	2697.71	5.46	10.1
Residual (po,e)	22783.25	978	23.30	23.30	43.3
Total	76387.87	1469			100

Source: UEW internship scores (2019)

From Table 23, the estimated variance component for interns (p), is 25.10 or 46.6% of the entire variance. The variance component for occasions accounts for 5.46 or 10.1% of the total variance. The second largest estimated variance component is the residual which is 23.30 or 43.3% of the entire variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. Hence, the major source of error for Technical Education for the 2016/2017 academic year is the p x o interaction and unidentified or random sources. The estimated variance component of the occasion facet forms only about one-tenth of the total variance.

Table 24 gives the ANOVA table of estimates of variance components for Vocational Education for 2016/2017 academic year.

Table 24 - ANOVA Estimates of Variance Components for Vocational Education for 2016/2017 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	9129.86	108	84.54	23.69	58.9
Occasions (o)	690.98	2	345.49	3.05	7.6
Residual (po,e)	2907.02	216	13.46	13.46	33.5
Total	12727.86	326			100

Source: UEW internship scores (2019)

It could be seen from Table 24 that, the estimated variance component for interns (p), 23.69, contributes as much as 58.9% to the entire variance. Though, the largest variance component, it is not a source of measurement error. The estimated variance component for occasions accounts for only 3.05 or 7.6% of the total variance. The second largest estimated variance component is that of the residual which is 13.46 or 33.5% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. Hence, major source of error in Vocational Education for the 2016/2017 academic year is the p x o interaction and unidentified or random sources. The occasion facet forms a little less than one-tenth of the total variance.

It could be concluded that for the mentors' results of the UEW-SIP for the 2016/2017 academic year, putting the variance component for interns (p) aside, the major source of error is the p x o interaction and unidentified or random sources and is followed by the occasion facet.

Table 25 gives the ANOVA table of estimates of variance components for Applied Science for 2017/2018 academic year.

Table 25 - ANOVA Estimates of Variance Components for Applied Science for 2017/2018 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	36792.34	341	107.90	25.82	45.0
Occasions (o)	807.92	2	403.96	1.09	1.9
Residual (po,e)	207.52	682	30.43	30.43	53.1
Total	58352.34	1025			100

Source: UEW internship scores (2019)

From Table 25, the estimated variance component for interns (p), 25.82, accounts for 45.0% of the whole variance. The estimated variance component for occasions accounts for only 1.09 or 1.9% of the total variance. The largest variance component is the residual which is 30.43 or 53.1% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. Hence, the major source of error in Applied Science for the 2017/2018 academic year is the p x o interaction and unidentified or random sources. The occasion facet forms only about one-fiftieth of the total variance.

Table 26 gives the ANOVA table of estimates of variance components for Business Education for 2017/2018 academic year.

Table 26 - ANOVA Estimates of Variance Components for Business Education for 2017/2018 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	36701.41	537	68.35	16.73	45.4
Occasions (o)	2118.03	2	1059.01	1.94	5.3
Residual (po,e)	19509.97	1074	18.17	18.17	49.3
Total	58329.41	1613			100

Source: UEW internship scores (2019)

Table 26 shows that, the estimated variance component for interns, 16.73, accounts for 45.4% of the total variance. The estimated variance component for occasions accounts for 1.94 or 5.3% of the total variance. The largest variance component is the residual which is 18.17 or 49.3% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. Hence, the major source of error in Business Education for the 2017/2018 academic year is the p x o interaction and unidentified or random sources. The occasion facet forms only about one-twentieth of the total variance.

Table 27 gives the ANOVA table of estimates of variance components for English and Communication for 2017/2018 academic year.

Table 27 - ANOVA Estimates of Variance Components for English and Communication for 2017/2018 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	8337.03	149	55.95	14.13	47.6
Occasions (o)	627.32	2	313.66	2.00	6.7
Residual (po,e)	4034.68	298	13.54	13.54	45.6
Total	12999.03	449			100

Source: UEW internship scores (2019)

From Table 27, the estimated variance component for interns (p), 14.13, forms the largest proportion of total variance which is 47.6%. The estimated variance component for occasions accounts for 2.00 or 6.7% of the total variance. The second largest contributor to total variance is the residual which is 13.54 or 45.6%. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error, therefore, in English and Communication for the 2017/2018 academic year is the p x o interaction and unidentified or random sources. The occasion facet contributes only a little less than one-tenth to total variance.

Table 28 gives the ANOVA table of estimates of variance components for Foreign Languages and Linguistics for 2017/2018 academic year.

Table 28 - ANOVA Estimates of Variance Components for Foreign Language and Linguistics for 2017/2018 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	43396.05	352	123.28	31.58	52.4
Occasions (o)	186.72	2	92.86	0.18	0.3
Residual (po,e)	20097.62	704	28.55	28.55	47.3
Total	63679.39	1058			100

Source**:** UEW internship scores (2019)

From Table 28, the estimated variance component for interns (p), 31.58, contributes as much as 52.4% to the total variance. The estimated variance component for occasions (o) accounts for only 0.18 or 0.3% of the total variance. The second largest variance component is the residual which is 28.55 or 47.3% of the entire variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error, therefore, in Foreign Languages and Linguistics for the 2017/2018 academic year is the p x o interaction and unidentified or random sources. The occasion facet contributes less than one-hundredth (0.3%) to total variance.

Table 29 gives the ANOVA table of estimates of variance components for Natural Science for 2017/2018 academic year.

Table 29 - ANOVA Estimates of Variance Components for Natural Science for 2017/2018 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	56100.18	413	135.84	36.69	58.1
Occasions (o)	590.58	2	295.29	0.65	1.0
Residual (po,e)	21294.76	826	25.78	25.78	40.8
Total	77985.51	1241			100

Source**:** UEW internship scores (2019)

From Table 29, the estimated variance component for interns (p), 36.69, accounts for the largest proportion of 58.1% of the total variance. The estimated variance component for occasions (o) accounts for only 0.65 or 1.0% of the total variance. The second largest estimated variance component is the residual which is 25.78 or 40.8% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error in Natural Science for the 2017/2018 academic year is the p x o interaction and unidentified or random sources. The occasion facet contributes only 1.0% of variability to total variance.

Table 30 gives the ANOVA table of estimates of variance components for Social Science for 2017/2018 academic year.

Table 30 - ANOVA Estimates of Variance Components for Social Science for 2017/2018 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	113497.28	1273	89.16	20.03	40.7
Occasions (o)	548.94	2	274.47	0.19	0.4
Residuals (po,e)	73967.73	2546	29.05	29.05	59.0
Total	188013.94	3821			100

Source: UEW internship scores (2019)

From Table 30, the estimated variance component for interns (p), 20.03, accounts for 40.7% of the total variance. The estimated variance component for occasions (o) accounts for only 0.19 or 0.4% of the total variance. The largest variance component is the residual which is 29.05 or 59.0% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error, therefore, in Social Science for the 2017/2018 academic year is the p x o interaction and unidentified or random sources. The occasion facet contributes so smaller a percentage (0.4%) to total variance.

Table 31 gives the ANOVA table of estimates of variance components for Technical Education for 2015/2016 academic year.

Table 31 - ANOVA Estimates of Variance Components for Technical Education for 2015/2017 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	14707.65	273	53.87	13.72	45.1
Occasions (o)	2184.59	2	1092.30	3.94	13.0
Residuals (po,e)	6950.07	546	12.73	12.73	41.9
Total	23842.32	821			100

Source**:** UEW internship scores (2019)

Table 31 shows that, the estimated variance component for interns (p), 13.72, forms as much as 45.1% of the entire variance. The estimated variance component for occasions (o) accounts for 3.94 or 13% of the total variance. The second largest variance component is the residual which is 12.73 or 41.9% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. The major source of error, therefore, in Technical Education for the 2017/2018 academic year is the p x o interaction and unidentified or random sources. The occasion facet forms a little above one-tenth of the total variance.

Table 32 gives the ANOVA table of estimates of variance components for Vocational Education for 2017/2018 academic year.

Table 32 - ANOVA Estimates of Variance Components for Vocational Education for 2017/2018 (Design: p x o)

Source of Variation	Sum of Squares	Df	Mean Squares	Variance Component	% of Total Variance
Interns (p)	5412.70	85	63.68	15.55	43.8
Occasions (o)	541.54	2	270.77	2.95	8.3
Residual (po,e)	2895.79	170	17.03	17.03	47.9
Total	8850.03	257			100

Source: UEW internship scores (2019)

From Table 32, the estimated variance component for interns (p), 15.55, accounts for 43.8% of the total variance. The estimated variance component for occasions (o) accounts for 2.94 or 8.3% of the total variance. The largest variance component is the residual which is 17.03 or 47.9% of the total variance. The two variance components, that is, for occasions and the residual contribute to measurement errors. Hence, the major source of error in Vocational Education for the 2017/2018 academic year is the p x o interaction and unidentified or random sources. The occasion facet forms about nearly one-tenth of the total variance.

It could be concluded that, for the mentors' results of the UEW-SIP for the 2017/2018 academic year, the major source of error is the p x o interaction and unidentified or random sources and is followed by the occasion facet.

It can therefore be concluded that for each academic year from 2015/2016 to 2017/2018, the major source of error in the mentors' results of the UEW-SIP is the p x o interaction followed by the occasion facet. It is also seen from the 24 ANOVA tables of analyses that in 13 of them, the universe score variance accounts for most of the variability in the observed score variance. In 11 of them, the universe score variance is the second largest. In all, the percentage range of the universe score variance is from 32.6% for Technical Education in 2015/2016 to 58.9% for Vocational Education and Natural Science in the 2016/2017 academic year. The occasion facet explains the least variability in all thematic course areas. According to Huigen et al. (2016), “a reliable instrument should have a high proportion of the variance explained by differences between the observed teachers and a low proportion of the variance explained by lessons and observers” (p.170). In conclusion, to a greater extent, the measurement instrument (i.e., the ITEF) of the UEW-SIP is reliable in the evaluation of teaching skills (Brennan, 2010; Shavelson and Webb, 1991).

Research Question 3

What is the optimum number of occasions of rating needed to obtain dependable mentors’ scores in the UEW-SIP for each academic year from 2015/2016 to 2017/2018?

Research Question 3 sought to find the ideal number of occasions of rating needed to obtain more dependable scores in the UEW-SIP for the academic years from 2015/2016 to 2017/2018.

Table 33 gives six different numbers of occasion and their G coefficients in a D study for relative ($E\rho^{2*}$) and absolute (Φ^{*}) interpretations for the 2015/2016 academic year. Appendix E_1 gives the output of the D study. The asterisks (*) indicates D study coefficients.

Table 33 – G Coefficients of D Study for Numbers of Occasion for Mentors' Results for 2015/2016 Academic Year

	Numbers of Occasion											
	1		2		3		4		5		6	
Specialism	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}
Applied Science	.56	.55	.66	.66	.74	.74	.79	.79	.83	.83	.85	.85
Business	.52	.44	.69	.61	.77	.70	.81	.76	.85	.80	.87	.83
English and Communication	.53	.37	.69	.54	.77	.64	.82	.70	.85	.74	.87	.78
Foreign Languages and Linguistics	.50	.50	.66	.66	.75	.75	.80	.80	.83	.83	.86	.85
Natural Science	.55	.55	.71	.71	.78	.78	.83	.83	.86	.86	.88	.88
Social Science	.41	.40	.58	.57	.67	.67	.73	.73	.77	.77	.80	.80
Technical	.43	.33	.60	.49	.69	.59	.75	.66	.79	.71	.81	.74
Vocational	.50	.46	.67	.63	.75	.72	.80	.77	.83	.81	.86	.84

Source: UEW internship scores (2019)

From Table 33, for a minimum of four occasions of rating, five thematic course areas: Business, English and Communication, Foreign Languages and Linguistics, Natural Science and Vocational, attained Coef_Grelative ($E\rho^{2*}$) of at least 0.80. For these thematic course areas, only two, which are Foreign Languages and Linguistics and Natural Science attained Coef_G absolute (Φ^{*}) of at least 0.80 and therefore these five thematic course areas failed to meet the set standard. Again, for a minimum of five occasions of rating, five thematic course areas, Applied Science, Business, Foreign Languages and Linguistics, Natural Science and Vocational attained both Coef_G absolute (Φ^*) and Coef_G relative ($E\rho^{2*}$) of at least 0.80. It can therefore be concluded that generally, for the 2015/2016 academic year in the UEW-SIP, the optimum number of occasions for both reliable and dependable results is five.

Table 34 gives six different numbers of occasions and their G coefficients in a D study for relative ($E\rho^{2*}$) and absolute (Φ^*) interpretations for the 2016/2017 academic year. Appendix E_2 gives the output of the D study.

Table 34 - G Coefficients of D Study for Numbers of Occasion for Mentors' Results for 2016/2017 Academic Year

	Numbers of Occasion											
	1		2		3		4		5		6	
Specialism	$E\rho^{2*}$	Φ^*	$E\rho^{2*}$	Φ^*	$E\rho^{2*}$	Φ^*	$E\rho^{2*}$	Φ^*	$E\rho^{2*}$	Φ^*	$E\rho^{2*}$	Φ^*
Applied Science	.39	.39	.57	.56	.66	.66	.72	.72	.77	.76	.80	.79
Business	.59	.54	.74	.70	.81	.78	.85	.82	.88	.85	.89	.87
English and Communication	.47	.45	.64	.62	.73	.71	.78	.76	.81	.80	.84	.83
Foreign Languages	.59	.58	.74	.74	.81	.81	.85	.85	.88	.87	.89	.89
Natural Science	.60	.59	.75	.74	.82	.81	.86	.85	.88	.88	.90	.90
Social Science	.42	.42	.59	.59	.69	.68	.74	.74	.78	.78	.81	.81
Technical	.52	.45	.68	.62	.76	.71	.81	.77	.84	.80	.86	.83
Vocational	.64	.59	.78	.74	.84	.81	.88	.85	.90	.88	.91	.90

Source: UEW internship scores (2019)

From Table 34, for a minimum of four occasions of rating, five thematic course areas: Business, Foreign Languages and Linguistics, Natural Science, Technical and Vocational, attained Coef_G relative ($E\rho^{2*}$) of at least 0.80. For these five thematic course areas, all, with the exception of Technical, attained Coef_G absolute (Φ^{*}) of at least 0.80 and therefore they failed to meet the set standard. Also, for a minimum of five occasions of rating, six thematic course areas, Business, English and Communication, Foreign Languages and Linguistics, Natural Science, Technical and Vocational attained both Coef_G absolute (Φ^{*}) and Coef_G relative ($E\rho^{2*}$) of at least 0.80. It can therefore be concluded that generally, for the mentors' results of the 2016/2017 academic year in the UEW-SIP, the optimum number of occasions for both reliable and dependable results is five.

Table 35 gives six different numbers of occasions and their G coefficients (relative [$E\rho^{2*}$] and absolute [Φ^{*}]) in a D study for the 2017/2018 academic year.

Table 35 - G coefficients of D Study for Numbers of Occasion for Mentors' Results for 2017/2018 Academic Year

	Numbers of Occasion											
	1		2		3		4		5		6	
Specialism	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}
Applied Science	.46	.45	.63	.62	.72	.71	.77	.77	.81	.80	.84	.83
Business	.48	.45	.65	.62	.73	.71	.79	.77	.82	.81	.85	.83
English and Communication	.51	.48	.68	.65	.76	.73	.81	.78	.84	.82	.86	.85
Foreign Languages	.53	.52	.69	.69	.77	.77	.82	.81	.85	.85	.86	.86
Natural Science	.59	.58	.74	.74	.81	.81	.85	.85	.88	.87	.90	.89
Social Science	.41	.41	.58	.58	.67	.67	.73	.73	.78	.77	.81	.80
Technical	.52	.45	.68	.62	.76	.71	.81	.77	.84	.80	.87	.83
Vocational	.48	.44	.65	.61	.73	.70	.78	.76	.82	.80	.85	.82

Source: UEW internship scores (2019)

From Table 35, for a minimum of five occasions of rating, all the thematic course areas with the exception of Social Science attained both Coef_G relative ($E\rho^{2*}$) and Coef_G absolute (Φ^{*}) of at least 0.80. It can therefore be concluded that generally, for the 2017/2018 academic year in the UEW-SIP, for mentors' results, the optimum number of occasions for both reliable and dependable results is five.

Research Question 4

What is the optimum number of occasions of rating needed to obtain dependable mentors' results in the UEW-SIP?

Research question 4 sought to find out optimum number of occasions of rating needed to obtain both reliable and dependable mentors' results in the UEW-SIP. The optimum numbers of occasions for reliable and dependable mentors' results in the UEW-SIP for the three academic years from Tables 35, 36 and 37 are as follows:

For 2015/2016 Academic Year: The optimum number of occasions for both reliable and dependable results is five.

For 2016/2017 Academic Year: The optimum number of occasions for both reliable dependable results is five.

For 2017/2018 Academic Year: The optimum number of occasions for both reliable and dependable results is five.

The three academic years give the same results that, the optimum number of occasions for both reliable and dependable results should be five. This is then compared to the D study results from the combined scores of the three academic years.

Table 36 gives the results of the D study from the combination of the scores of the three academic years from 2015/2016 to 2017/2018. Appendix E_4 shows the output of the D study.

Table 36 - G coefficients of D Study for Numbers of Occasions for Mentors' Results from 2015/2016 to 2017/2018 Academic Years

Numbers of Occasion											
1		2		3		4		5		6	
$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}	$E\rho^{2*}$	Φ^{*}
.52	.51	.68	.68	.77	.76	.81	.81	.84	.84	.87	.86

Source: UEW internship scores (2019)

From Table 36, for a minimum of four occasions of rating, both Coef_G relative ($E\rho^{2*}$) and Coef_G absolute (Φ^{*}) reached the threshold of 0.80 (i.e., 0.81 each). For five occasions of rating, both Coef_G relative ($E\rho^{2*}$) and Coef_G absolute (Φ^{*}) again reached the threshold of 0.80 (i. e., 0.84 each). Coef_G relative ($E\rho^{2*}$) and Coef_G absolute (Φ^{*}) of 0.84 for five occasions of rating is much more above the threshold of 0.80 and therefore more stable than the Coef_G relative ($E\rho^{2*}$) and Coef_G absolute (Φ^{*}) of 0.81 for four occasions. Five occasions of rating is therefore accepted as the optimum number of occasions for more stable and dependable results in consonance with the conclusions from the three separate academic years.

It can therefore be concluded that, for mentors' results in the UEW-SIP, to ensure economy of use of resources, the optimum number of occasions for reliable and dependable results should be five.

Discussion of Research Findings

In this section, the findings of the study are discussed in relation to:

i. Reliability and dependability of the mentors' results of the UEW-SIP.

ii. Measurement errors in the mentors' results of the UEW-SIP.

iii. Optimum number of occasions for reliable and dependable mentors' results for each academic year in the UEW-SIP.

iv. Optimum number of occasions for reliable and dependable results for the UEW-SIP.

Reliability and dependability of the mentors' results of the UEW-SIP

The first finding of the study is that for the 2015/2016 to 2017/2018 academic years, for relative interpretation ($E\rho^2$), the mentors' results of the UEW-SIP were strongly reliable (Coef_G relative [$E\rho^2$] of 0.66 – 0.84). For absolute interpretation (Φ), the results were moderately to strongly dependable (Coef_G absolute [Φ] of 0.59 – 0.81). For absolute interpretation using the Phi(lambda) coefficient, the results were strongly dependable ($\Phi[\lambda]$ of 0.98 – 0.99).

As stated by Shavelson and Webb (1991), the G coefficient for relative interpretation ($E\rho^2$) reflects the proportion of variability in individuals' scores that is systematic or attributable to universe score variability. Therefore, the Coef_G relative ($E\rho^2$) of 0.66 to 0.84 reflects the proportion of obtained score variance attributable to systematic differences in interns' knowledge of subject matter and skills in teaching. It is their universe-score variability. It also gives a practical index of the quality of the measurement design vis-à-vis the measurement instrument of the UEW-SIP on a scale of 0.0 - 1.0 (Marcoulides, 2000) and so it can be concluded that the UEW-SIP measurement design was high in quality.

The Coef_G absolute (Φ) of 0.59 – 0.81 is an index of the dependability of the results emanating from the UEW-SIP measurement procedure. It reflects the accuracy of generalising from an intern's observed score on an occasion to the average score that the intern would have received under all the possible occasions of lesson delivery (Feldt & Brennan, 1989; Shavelson & Webb, 1991; Haertel, 2006; Brennan, 2010). On a scale of 0.0 - 1.0, it could be said that the results are dependable to a greater extent.

It must be understood that the cut-off score for the UEW-SIP for all the years is 50.0 (i.e., minimum pass mark in UEW). For the academic years used in the study, all the computed sample means were greater than this cut-off score. Therefore, the difference between the sample mean for each academic year and the cut-off score is not a null value and so should be taken as a source of true variance with the sample mean being subject to sampling fluctuation (EDUCAN Inc. & IRDP, 2010). Hence, the most appropriate index for description of dependability in this context is Brennan and Kane's (1977) Phi(lambda) coefficient. The Phi(lambda) coefficient in this study ranges from 0.98 to 0.99 indicating strongest dependability resulting from an increment in the estimate of the true variance as a function of the distance from the sample mean to the cut-off score.

The finding above is consistent with the study by Burns and Froman (as cited in Burns, 1998), which aimed at determining the optimum conditions (number of items and occasions) for the reliable application of a modified form of the Habitual Physical Activity Index (HPAI) for use among older American adults. The HPAI was made up of an eight-item index for work physical activity and an eight-item index for leisure physical activity.

G-coefficients for both relative and absolute decisions were computed for the study. For the instrument's original eight items and two occasions, G coefficients for relative ($E\rho^2$) decisions for the work and leisure indices were 0.86 and 0.80, respectively. G-coefficients for absolute (Φ) decisions for the work and leisure indices were 0.79 and 0.75, respectively. These conclusions on G-coefficients are similar to that of the present study with the only point of partial inconsistency being the G-coefficients for absolute decisions. Whereas in the study by Burns and Froman, for absolute interpretation, the results were strongly dependable, in the present study, the results are moderately to strongly dependable (Coef_G absolute [Φ] of 0.59 – 0.81). However, for absolute interpretation using the Phi(lambda) coefficient (Φ[λ] of 0.98 – 0.99), the findings are consistent.

This finding is again consistent with the finding of a study by Patrick et al. (2020), which aimed at evaluating the score stability of the Framework for Teaching (FFT), a prominent observation instrument used for teacher evaluation. The FFT's Classroom Environment, Instruction, and Total scores were decomposed into potential sources of variation (teachers, lessons, raters, and their interactions). Computed G coefficients for relative interpretation ($E\rho^2$) ranged from 0.92 to 0.96 for classroom environment and total scores, and they were 0.87 and 0.79 for reading and mathematics instruction respectively. These values are indicators of the higher quality (score stability) of the FFT just as seen in the present study with the measurement instrument of the UEW-SIP.

The finding of another study with which the finding of the present study is consistent, is that of Lakes and Hoyt (2009) on Child and Adolescent Psychology, which aimed at helping readers to understand the effects of measurement error on findings in clinical child and adolescent research. Five trained observers rated 181 elementary school children on three multi-item scales designed to measure different domains of self-regulation using the cognitive (6 items), affective (7 items) and physical (3 items) self-regulation scales. G coefficients for relative interpretation computed were 0.86 for cognitive, 0.92 for affective, and 0.88 for physical. Again, these values are indicators of the higher quality (score stability) of the self-regulation subscales just as seen in the present study with the measurement instrument of the UEW-SIP.

Measurement errors in the mentors' results of the UEW-SIP

The second finding of the study is that, the major source of error in the mentors' results of the UEW-SIP is the p x o interaction followed by the occasion facet for each academic year. Also, to a greater extent, the ITEF rating scale is reliable. The error variance is distributed among the academic years as follows. In the 2015/2016 academic year, error from the occasion facet ranged from as low as 0.1% for Natural Science to 29.9% for English and Communication. For the residual, the margin of error ranged from 33.3% for English and Communication to 58.6% for Social Science. In the 2016/2017 academic year, error from the occasion facet ranged from 0.4% for Foreign Languages and Linguistics to 10.1% for Technical Education. For error due to the residual, the range is 33.5% for Vocational Educational to 60.1% for Applied Science. Lastly, for the 2017/2018 academic year, error due to the occasion facet ranged from 0.3% for Foreign Languages and Linguistics to 13.0% for Technical Education. With the residual, the range of error is from 40.8% for Natural Science to 59.0% for Social Science.

Error from the occasion facet is explained as variability in interns' scores that results from differential occasional conditions. Different occasions would present different instructional periods and circumstances of lesson presentation and assessment and these would cause generalisation from the occasion sample to the occasion universe to be less accurate (Shavelson & Webb, 1991). This error of generalisation from the occasion facet is what is represented as 0.1% to 29.9% in the 2015/2016 academic year alone.

Error variance from the p x o interaction is explained as the variability that results from inconsistencies from one occasion to another in particular intern's behaviour (Shavelson & Webb, 1991; Brennan, 2010). For instance, some interns may perform better when they teach at certain times of the day, some may have emotional disturbances on some occasions, and lessons taken at the later end of the internship period may also suffer from fatigue on the part of interns. So, the match between an intern's personal peculiarities and a particular occasion constitutes an interaction between interns and occasions which results in inconsistencies from one occasion to another in particular person's behaviour. This increases variability and causes generalisation from a student's score on an occasion to his average score over all possible occasions in the occasion universe—the universe score, to be less accurate.

In this study, the proportion of error from the p x o interaction is quite huge, for example, 33.5% to 60.1% for 2016/2017 academic year. The obvious reason is that this study involves only the occasion facet with any other probable facets such as the rater and type of lesson taught, treated as unmeasured facets in the study. These are obvious sources of non-sampled random fluctuations at play in the study to swell up the error margin.

The argument above is given credence by Shavelson and Webb (1991), Brennan (2010) and Cardinet et al. (2010) that after accounting for the error due the occasion facet, it cannot be known exactly whether further differences in occasion scores reflect the p x o interaction or random unidentified sources of variability. Hence, these two sources of variability are put together as a residual and defined by the *p* × *o* interaction confounded by other sources of variability.

The findings above are consistent with the findings of a study by Huijgen et al. (2016) which aimed at developing a consistent observation instrument (FAT-HC) and scoring design to assess the means by which history teachers promote historical contextualisation in their classrooms. The consistency in findings is seen in terms of the instrument's reliability and major source of error. With the facets being history teachers (t), history lessons (l) and observers (o) in a fully crossed design, the teacher facet explained the largest proportion of variability (59.1%) in the observed scores which indicated a high reliability of the measurement instrument. This was followed by the residual (34.7%). This is consistent with 13 of the 24 G study ANOVA analyses in the present study where the universe score variance was the largest followed by the residual.

The findings are also consistent with a study by Ramadan et al. (2019). The aim of the study was to make a standard instrument to assess the competencies of physics teachers in social, pedagogic, professional and personality skills. The study used a two-facet nested design, *p x (i: r)*, with teachers (p), items (i) and raters (r). The consistency in the findings is the fact that the major source of error was the residual (52.3%). Also, the variability accounted for by the object of measurement (9.2%) agreed with the 11 cases of G study analysis in the present study in which the universe score variance is the second largest source of variability in the observed score variance. This has a restrictive effect on the reliability of the measurement instrument (Shavelson & Webb, 1991; Brennan, 2010).

The findings are however, not consistent with the findings of the Burns and Froman's study (as cited in Burns, 1998) on the reliable application of a modified form of the HPAI for use among older American adults. The estimated variance components that resulted from the G study analysis of the fully crossed $s \times i \times o$ design showed that the universe score explained the second largest proportion of variability (24%) in the observed scores which was followed by the residual (22%), with rather the $s \times i$ interaction accounting for the largest proportion (i.e., major source of error with 37%). So here, the residual is not the major source of error.

The explanation for the case above is that it is a three-facet crossed study that untangles the error of measurement and spreads it to as many sources as possible (Brennan, 2010). This results in a relatively smaller variance from the residual. Once the universe score variance does not explain most of the variability in the observed score variance, there is a restrictive effect on the reliability of the HPAI instrument (Shavelson and Webb, 1991; Brennan, 2010).

Optimum number of occasions for reliable and dependable mentors' results for each academic year in the UEW-SIP

The third finding of the study is that generally, for the mentors' results in the UEW-SIP from 2015/2016 to 2017/2018 academic years, the optimum number of occasions for reliable and dependable results, for each academic year, is five.

Taking the findings one academic year at a time, for 2015/2016, for three occasions of rating which is the practice in the UEW-SIP, the highest G coefficients were Coef_G relative ($E\rho^2$) of 0.78 and Coef_G absolute (Φ) of 0.78 for Natural Science. It was only Natural Science that came quite close to the benchmark of 0.80. Some thematic course areas had as low G coefficients as Coef_G relative ($E\rho^2$) of 0.67 and Coef_G absolute (Φ) of 0.67 for Social Studies while Technical Education had Coef_G relative e ($E\rho^2$) of 0.69 and Coef_G absolute (Φ) of 0.59. These values were deemed unsatisfactory because they fell below the G coefficient benchmark of 0.80.

For 2016/2017, for the standard three occasions of rating, four thematic course areas met the benchmark of G coefficient of 0.80 with only one of the four (i.e., Business Education) having a lower Coef_G absolute (Φ) of 0.78. Since not more than half of the thematic course areas attained the benchmark G coefficient of 0.80, the situation was deemed unsatisfactory.

For 2017/2018, for the standard three occasions of rating, only Natural Science Education met the standard G coefficient of 0.80 for both relative and absolute interpretations. Social Science attained as low G coefficient as 0.67 for both Coef_G relative ($E\rho^2$) and Coef_G absolute (Φ). These values were also deemed unsatisfactory.

The results above necessitated a D (optimization) study to be conducted for each academic year with varied numbers of occasion from one to six in order to obtain an optimum number of occasions at which more than half of the thematic course areas would attain standard G coefficients of at least 0.80. For the numbers of occasion that would be arrived at for each academic year, we shall be assured of highly stable and dependable universe (true) scores upon which generalisation from sample to the universe of admissible observations could be made with utmost accuracy.

This third finding with its approach is generally consistent with the main purpose of most G theory studies and fully utilises the major strength of G theory over CTT. In the assertions of Shavelson and Webb (1991), Brennan (2010), Li et al. (2015) and Etsey (2015), G theory gives a framework that can be used to pin-point and quantify the sources of error on which decisions can be made to optimise the measurement procedures so as to give more reliable and dependable scores.

The approach, purpose and the finding of the present study are consistent with those of Lakes and Hoyt (2009) on Child and Adolescent Psychology. The conclusion from their study was that researchers assessing cognitive self-regulation among children must measure across occasions and not just once in order to achieve stable and dependable results. This was after a D study had shown that using a design with 10 raters instead of the original five and the RCS in its current form (7, 6, and 3 items for the Cognitive, Affective, and Physical subscales, respectively) on a single occasion, a researcher will obtain G coefficients for relative interpretation of 0.47 for Cognitive, 0.83 for Affective, and 0.76 for Physical subscales. But increasing the number of testing occasions to four increases the expected G coefficients for relative interpretation to 0.71 for Cognitive, 0.90 for Affective, and 0.85 for Physical subscales.

Again, this finding is consistent with the finding of Froman et al. (as cited in Burns, 1998) which was on the usefulness of generalizability coefficients in determining stability and dependability of results. The study aimed at assessing the generalizability of people's attitudes toward Persons With AIDS (PWA) using the AIDS Attitude Scale (AAS), which comprises the empathy and avoidance scales, for relative and absolute interpretations. The conclusions from a D study were that the empathy subscale had acceptable G coefficients under both relative and absolute decisions across one or two observations, while the avoidance subscale had to be administered for more than once, especially if the scores are meant for an absolute decision. The original design of this measurement procedure was a two-facet item (i) by occasion (o) crossed design which was administered on two occasions.

Finally, the approach, purpose and this finding of the present study are consistent with those of Patrick et al. (2020). In evaluating the score stability of the Framework for Teaching (FFT), a D study was conducted and the results were that, two raters, each scoring three reading lessons or four mathematics lessons, were necessary to achieve sufficiently reliable total scores. For instruction scores, three raters each scoring seven reading lessons were needed, more than four raters each scoring eight lessons were needed for mathematics. The original measurement design of the FFT was three raters each scoring 200 reading and mathematics lessons taught by 20 kindergarten teachers.

Optimum number of occasions for reliable and dependable results for the UEW-SIP

The fourth finding of this study is that, the optimum number of occasions for both reliable and dependable results should be five to ensure economy of use of resources in the mentors' assessment of the UEW-SIP. This is the general finding that climaxes the study and lays the foundation for the improvement of the mentors' assessment of the UEW-SIP that has not been evaluated in terms of the psychometric properties of the results since its inception.

The approach, purpose and the finding of this present study are consistent with that of the study by Ramadan et al. (2019). A D study conducted showed that to reach a G coefficient for relative interpretation ($E\rho^2$) of at least 0.70 which is acceptable for research purposes (Brennan & Kane, 1977), the assessor must increase the items to four (i.e., use indicators 1, 2, 3, and 4). They concluded that, to get the results of an assessment of genuine physics teacher competencies, the instrument developed can be used with four competency indicators.

Again, the approach, purpose and the finding of the present study are consistent with those of Burns and Froman (as cited in Burns, 1998), which aimed at determining the optimum conditions (number of items and occasions) for the reliable application of a modified form of the HPAI. At a benchmark standard G coefficient of 0.80 and one administration of the HPAI scale to save cost, the following results were obtained that led the researchers to modify the HPAI scale. For relative decisions, for the work index, a G-coefficient ($E\rho^2$) of 0.80 was obtained using eight items and one occasion. For relative decisions for the leisure index, 10 items were needed to bring the G-coefficient ($E\rho^2$) to greater than 0.80. Concerning absolute decisions, with the work index, 11 items on one occasion were needed to achieve the 0.80 criterion. For the leisure index, 13 items on one occasion were needed to bring the D-coefficient (Φ) to 0.80.

Finally, the purpose and the finding of the present study are consistent with those of Huijgen et al. (2016) which aimed at developing the FAT-HC instrument and scoring design. These researchers were interested in research purposes and formative evaluations and so focused on the absolute level of individuals' performance irrespective of others' performance. Hence, they used the index of dependability coefficient (Φ) to find the optimum number of observers. With a benchmark phi coefficient (Φ) of at least 0.80, it was found that the optimum scoring design would use two observers to evaluate two different lessons taught by the same teacher (Φ = 0.83) or three observers to evaluate the same lesson taught by one teacher (Φ = 0.80).

The three studies reviewed in the three paragraphs above had one main objective and that was to use G theory to modify an existing measurement procedure to make it more efficient. This is the major aim of the present study. This was emphasised strongly by Cronbach et al. (as cited in Lakes & Hoyt, 2009) as far back as 1972 that, researchers developing new measurement procedures should first carry out a G study, to help guide the design and interpretation of later D studies to help come out with the most reliable measurement procedures.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Overview

The main purpose of this study was to determine the dependability of the mentors' results of the UEW-SIP, using G theory. Inherent in this overall purpose were to find the reliability and the sources of error in the mentors' results. Data were analysed by performing a univariate generalizability analysis using EduG version 6.1. The documentary research design was used for the study with a random effect one-facet crossed design, in which interns (*p*) were crossed with occasions (*o*).

The study was based on G theory and was carried out in UEW, Ghana, using eight out of 14 academic faculties. A total of 9,082 bachelor's degree graduates' results for the academic years 2015/2016, 2016/2017 and 2017/2018 were used for the analysis.

Summary of Findings

The following are the main findings from the data analysis of the mentors' results of the UEW-SIP, for the 2015/2016 to 2017/2018 academic years.

For stability and dependability, for relative interpretation, the results were strongly reliable (Coef_G relative [$E\rho^2$] of 0.66 – 0.84). For absolute interpretation, the results were moderately to strongly dependable (Coef_G absolute [Φ] of 0.59 – 0.81). For absolute interpretation using Phi(lambda) coefficient, the results were strongly dependable ($\Phi(\lambda)$ of 0.98 – 0.99).

For major sources of error, the major source of error in the mentors' results is the p x o interaction followed by the occasion facet for each academic year. In addition, to a greater extent, the measurement instrument (i.e., the ITEF) of the UEW-SIP is reliable in the evaluation of teaching skills.

For optimum number of occasions for dependable results for each academic year, from 2015/2016 to 2017/2018 academic years, for both stable and dependable results, the number of occasions is five.

For optimum number of occasions for dependable results in the UEW-SIP, to ensure economy of use of resources, the number of occasions for both stable and dependable results should be five.

Conclusions

It could be concluded that, the mentors' results of the UEW-SIP for the period of 2015/2016 to 2017/2018 were generally strongly reliable indicating that over the three occasions that the skill of teaching is measured, it is highly stable. Also, in generalising the performance on a single occasion to obtain a universe (true) score in the universe of admissible occasions, a generally higher degree of accuracy is assured, especially when the Phi(lambda) coefficient is used.

For the major sources of error, since this study was one facet, it could not capture all the facets that could contribute to error in the results. It can therefore be concluded that the p x o interaction combined with other unmeasured facets form the major source of error in the results and this is followed by the occasion facet.

The optimum number of occasions for dependable results for each academic year in the UEW-SIP is concluded as five. At a minimum of five occasions that acceptable G coefficients for formative evaluation can be arrived at for both stable and dependable results.

The optimum number of occasions for dependable results for the entire UEW-SIP is concluded as five. At a minimum of five occasions, acceptable G coefficients that are beyond the threshold of 0.80 for formative evaluation can be arrived at for more stable and dependable results.

Recommendations

In view of the above research findings and the conclusions arrived at, the following recommendations are made.

1. The findings of the study establish the fact that the mentors' results of the UEW-SIP are generally strongly stable and dependable. This ensures trust and confidence in the use of the results. It is recommended that the UEW should officially document this important psychometric property of the scores in reference to the rating scale used for the measurement of teaching skills. This will help know the quality of the results and of the measurement procedure to boost confidence in their use and to aid planning and improvement of the internship programme.
2. Establishing in the findings that the mentors' results are generally strongly stable and dependable without knowing anything about the psychometric properties of the university lecturers' scores is not good enough for the UEW-SIP. This is because the mentors' scores form only a percentage of the total score for grading. It is recommended that the implementors (i.e., ITECPD of UEW) of the internship programme should increase the university supervisors' rating from one to at least

two occasions so that G theory can be applied to find the stability and dependability of the results over occasions of the internship period. This would give a more holistic picture of the quality of the UEW-SIP.

3. Based on the last finding of the study that the optimum number of occasions for stable and dependable mentors' results is five, it is recommended that the implementors (i.e., ITECPD of UEW) of the internship programme should increase the number of occasions for mentors' rating from three to five. This calls for more commitment on the part of school mentors as this will undoubtedly increase their workload.
4. The mentors' results of the UEW-SIP for one academic year (2015/2016 for Technical Education) were found to be moderately to strongly dependable. This case of moderately dependable scores is not an ideal situation for decision making. It is therefore recommended that, the developers of the ITEF should redesign it by specifying clearly the breakdowns of indicators of expected behaviours for the items on it and corresponding points to be awarded for such indicators. This would give more precise descriptions for scoring to ensure more consistent scores.

Implication of the Study for Educational Practice

The findings of the study establish the fact that, the mentors' results of the UEW-SIP are generally strongly stable and dependable, ensuring trust and confidence in the use of the results by decision makers. The implication drawn here for educational practice in the context of the UEW is that, the UEW-SIP has contributed positively to the training of teachers for the school system in

Ghana. This is, especially in the aspect of practical training. The UEW-SIP, which is largely practical work and on-the-job learning, should continue to be an integral part of the teacher training programme in UEW.

Suggestions for Further Research

The following are recommended for future research.

1. The study used only the occasion facet and so conclusions on measurement error were made only on the occasion facet. In order to have a complete picture of the major source(s) of error in the mentors' results of the UEW-SIP, it is suggested that this study is replicated using at least two facets to include the occasion (o) and the lesson taught (l).
2. The study used only the mentors' results of the UEW-SIP. The university raters' results could not be analysed due to the violation of a basic assumption in G theory application. In order to establish the psychometric properties (reliability and quality of the measurement design) of the entire UEW-SIP, it is suggested that the study is replicated with an additional application of CTT on the university raters' scores that are gathered on a single occasion.

REFERENCES

Adu-Gyamfi, S., Donkoh, W. J. & Addo, A. A. (2016). Educational reforms in Ghana: Past and present. *Journal of Education and Human Development*, *5*(3), 158-172.

Allen, M. J. & Yen, W. M. (1979). *Introduction to measurement theory.* Long Grove, Illinois: Waveland Press, Inc.

Amedahe, F. K. (2004). *Notes on educational research.* Unpublished document, University of Cape Coast, Ghana.

Anastasi, A. & Urbina, S. (2007). *Psychological testing.* (7th ed.). New Delhi: Prentice Hall.

Antwi, M. K. (1992). *Education: Society and development in Ghana.* Accra: Unimax Publishers Limited.

Asamoah-Gyimah, K. & Duodu, F. (2007). *Introduction to research methods in education.* Winneba: IEDE, UEW, Winneba, Ghana.

Atilgan, H. (2013). Sample size for estimation of G and Phi coefficients in generalizability theory. *Agitim Arastirmalari – Eurasian Journal of Research*, *51*, 215 – 228.

Baecke, J. A. H., Burema, J. & Frijters, E. R. (1982). A short questionnaire for the measurement of habitual physical activity in epidemiological studies. *American Journal of Clinical Nutrition. 36*(5), 936-942.

Becker, H. S. (1958). Problems of inference and proof in participant observation. *American Sociological Review, 23*(6), 652-660.

Brennan, B. L. (2010). *Generalizability theory.* New York: Springer.

Brennan, B. L. (2006). *Educational measurement*. New York: Praeger.

Brennan, R. L. (2001b). *Generalizability theory.* New York, NY: Springer.

Brennan, B. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice, 16*(4), 14 – 20.

Brennan, B. L. (1992). An NCME instructional module on generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27-34.

Brennan, B. L. & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*, 277 – 289.

Burns, K. J. (1998). Beyond classical reliability: Using generalizability theory to assess dependability. *Research in Nursing and Health*, *21*, 83 – 90.

Cardinet, J., Johnson, S. & Pini, G. (2010). *Generalizability theory using EduG*. New York: Routledge.

Centre for Teacher Development and Action Research (CETDAR). (2014). *Students' internship handbook.* (rev. ed.). CETDAR: University of Education, Winneba.

Crick, J. E. & Brennan, R. L. (1983). *Manual for GENOVA: A generalised analysis of variance system.* (ACT Technical Bulletin No. 43). Iowa City: IA: ACT, Inc.

Crocker, L. & Algina, J. (1986). *Introduction to classical & modern test theory*. Chicago: Holt, Rinehart & Winston, Inc.

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: Wiley.

Dobler, E., Kesner, C., Kramer, R., Resnik, M. & Devin, L. (2009). *Collaborative model for developing classroom management skills in urban professional development school settings.* (EJ915861). Retrieved from: http://eric.ed.gov /ERICWebPortal/home, on 11/06/18.

Dombrowski, S. C. (2015). *Psycho-educational assessment and report writing*. New York: Springer.

Dunn, D. S. (2001). *Statistics and data analysis for the behavioural sciences.* Boston, MA: McGraw-Hill.

EDUCAN Inc. & Institute for Research and Documentation in Pedagogy [IRDP]. (2010). *EduG user guide.* Retrieved from: http://www.irdp.ch/edumetrie/ englishprogram.htm, on 13/03/2019.

Educations.com. (2015). *Types of internships*. Retrieved from: https://www.educations.com/internships/types-of-internships, on 15/06/18.

Etsey, Y. K. A. (2015). The role of generalizability theory in the search for true scores in multiple measurements. *Journal of Educational Assessment in Africa, 10*, 79 – 91.

Faculty of Agriculture Education. (2013). *UEW General Agriculture re-accreditation document*. (Unpublished document). University of Education, Winneba.

Feldt, L. S. & Brennan, B. L. (1989). *Educational measurement.* (3rd ed.). New York, NY: McMillan Publishing Company.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver & Boyd.

Gall, M. D., Gall, J. P. & Borg, W. R. (2007). *Educational research: An introduction,* (8th ed.). London: Pearson.

Gates, S. M. & Pual, C. (2004). *Intern programmes as a human resource management tool for the Department of Defense.* Pittsburgh: Rand Corporation.

Gay, L. R., Mills, G. E. & Airasian, P. (2009). *Educational research: Competencies for analysis and applications.* (9th ed.). Upper Saddle River, New Jersey: Pearson.

Gugiu, M. R., Gugiu, P. C. & Baldus, R. (2012). Utilizing generalizability theory to investigate the reliability of grades assigned to undergraduate research papers. *Journal of Multidisciplinary Evaluation, 8*(19), 26-40.

Gulliksen, H. (1987). *Theory of mental tests.* London: Lawrence Erbaum Associates, Publishers.

Haertel, E. H. (2006). *Educational measurement.* (4th ed.). New York: Praeger Publishers.

Hoyt, W. & Melby, J. N. (1999). Dependability of measurement in counselling psychology: An introduction to generalizability theory. *The Counselling Psychologist, 27*(3), 321 – 352.

Huen, H. K. and Lei, P. (2007). Classical versus generalizability theory of measurement. *Educational Measurement, 4*, 1 – 13.

Huhman, H. R. (2011). How to talk to your student about different types of internships. *College Parents of America.* Retrieved from: http://collegeparents.org/2011/04/25, on 17/05/18.

Huijgen, T., Holthuis, P., Boxtel, T. & Grift, W. V. (2018). Promoting historical contextualization in classrooms: An observational study. *Educational Studies. 45*(3), 456 – 479.

Huijgen, T., Grift, W. V., Boxtel, T. & Holthuis, P. (2016). Teaching historical contextualisation: The construction of a reliable observation instrument. *European Journal of Psychology of Education.* Retrieved from: https://www.researchgate.net/publication/299477593, on 15/07/2020.

Hyman-Parker, S. (1998). Benefits and limitations of internships as viewed by educators and retailers. *ResearchGate.* Retrieved from: https://www.researchgate.net/publication/234600704, on 13/03/19.

Institute of Educational Development and Extension (IEDE). (2010). *Report of UEW Internship Planning Committee.* IEDE, UEW, Winneba, Ghana.

Kadingdi, S. (2006). Policy initiative for change and innovation in basic education programmes in Ghana. *ResearchGate*. Retrieved from: https://www.researchgate.net/publication/26598028, on 14/03/19.

Lakes, K. D. & Hoyt, W. T. (2009). Applications of generalizability theory to clinical child and adolescent psychology research. *Journal of Clinical Child and Adolescent Psychology, 38*(1), 144–165.

Li, M., Shavelson, R. J., Yin, Y. & Wiley, E. (2015). Generalizability theory. *International Collaboration for Performance Assessment of Learning in Higher Education: Research & Development* (iPAL: R&D). Retrieved from: DOI 10.1002/9781118625392.wbecp352, on 17/06/18.

Lindquist, E. F. (1953). *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin.

Loretto, P. (2017). Types of internship. *The balance.* Retrieved from: https://www.thebalance.com/types-of-internships-1986738, on 15/07/18.

Mahuron, S. (2019). *The cons of an internship*. Retrieved from: http://work.chron.com/cons-internship-3037.html, on 13/03/19.

Marcoulides, G. A. (2000). *Handbook of applied multivariate statistics and mathematical modeling*. Cambridge: Academic Press.

Masood, H. (2014). *Ten internship characteristics that attract exceptional interns.* Retrieved from: http://www.braatheenterprises.com/virtual project, on 13/05 /18.

McLeod, S. A. (2015). *Observation methods.* Retrieved from: www.simplypsychology.org/observation.html, on 12/02/18.

McWilliam, H. O. A. & Kwamena-Poh, M. A. (1975). *The development of education in Ghana.* London: Longman Group Ltd.

Miles, M. B. & Huberman, A. M. (1994). *Qualitative data analysis.* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Ministry of Education. (2016). *Teaching practice: Student teacher handbook. T- TEL schools partnership programme.* Retrieved from: http://www. t-tel.org., on 24/05/19.

National Association of Colleges and Employers [NACE]. (2018). *Building a premier internship programme: A practical guide for employers.* Retrieved from: https://www.naceweb.org/talent-acquisition/ internships/15-best-practices-for-internship, on 10/05/18.

National Board for Professional and Technician Examination [NABPTEX]. (2020). *Our core departments.* Retrieved from: https://www.nabptex. gov.gh/core-departments, on 23/06/20.

Nitko, A. J. (1996). *Educational assessment of students*. (3rd ed.). New Jersey: Prentice-Hall, Inc.

Odunga, J. (2014). Observation methods of data collection in behavioral science. *Technology*. Retrieved from: https://www.slideshare.net/ jakeodunga, on 17/06/18.

Oduro-Okyireh, G. (2013). Testing practices of senior high school teachers in the Ashanti Region of Ghana. *Journal of Counselling, Education and Psychology*, *3*(1), 76 – 87.

Ohio State University Department of Political Science. (2013). *Internship.* Retrieved from: https://polisci.osu.edu/sites/polisci.osu.edu, on 16/06/18.

Patrick, H., French, B. F. & Mantzicopolous, P. (2020). The reliability of framework for teaching scores in kindergarten. *Journal of Psychoeducational Assessment.* Retrieved from: https://doi.org/10.1177/0734282920910843, on 11/07/2020.

Patterson, V. (1997). *The employers' guide: Successful intern programmes.* (EJ550307). Retrieved from: https://eric.ed.gov/?id=EJ550307 on 15/06/18.

Putka, D. J. & McCloy, R. A. (2008). Estimating variance components in SPSS and SAS: An annotated reference guide. *Human Resources Research Organisation.* Retrieved from: http://www.humrro.org, on 15/05/18.

Ramadhan, S., Nasran, S. A., Utomo, H. B., Musyadad, F. & Ishak, S. (2019). The implementation of generalizability theory on Physics teachers' competency assessment instruments development. *International Journal of Scientific and Technology Research*, *8* (7), 332 – 337.

Rentz, J. O. (1987). Generalizability theory: A comprehensive method for assessing and improving the dependability of marketing measures. *Journal of Marketing Research, XXIV*, 19 – 28.

Sarantakos, S. (1993). *Social research.* (2nd ed.). New Delhi: Palgrave.

Shavelson, J. S. & Webb. N. M. (2005). *Generalizability theory.* Retrieved from: https://web.stanford.edu/dept/SUSE/SEAL/Reports, on 17/06/18.

Shavelson, J. S. & Webb. N. M. (1991). *Generalizability theory: A primer*. London: SAGE Publications.

Shavelson, J. S. & Webb, N. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34*, 133-166.

Stora, B., Hagtvet, K. A. & Heyerdahl, S. (2013). Reliability of observers' subjective impressions of families: A generalizability theory approach. *Psychotherapy Research.* Retrieved from: https://doi.org/10.1080/105 0330 7.2012.733830, on 27/06/18.

Technical Education Unit. (2010). *General overview of the operations of the TEU.*Retrieved from: https://www.gesteu.com/about.html, on 15/11/19.

Traub, R. E. (1997). *Classical test theory in historical perspective.* Retrieved from: www.winsteps.com/a/Traub.pdf, on 15/11/18.

University of St. Thomas, Minnesota. (2018). *Characteristics of a quality, world class internship.* Retrieved from: https://www.stthomas.edu/ career, on 12/04/18.

University of Education, Winneba. (2016). *UEW 20th congregation basic statistics.* Winneba: UEW Publication Unit.

University of Education, Winneba. (2017). *UEW 21st congregation basic statistics*. Winneba: UEW Publication Unit.

University of Education, Winneba. (2017). *UEW 22nd congregation basic statistics*. Winneba: UEW Publication Unit.

University of Education, Winneba. (2017). *UEW annual diary*. Winneba: UEW Publication Unit.

Walton, J. K. G. (2015). *Early Ghana Methodism.* Accra: Artegraphics.

Webb, N. M., Shavelson, R. J. & Haertel, E. H. (2006). *Handbook of statistics*, (Vol. 26). New York: Elsevier B. V.

Worthen, B. R. & Sanders, J. R. (1987). *Educational evaluation. Alternative approaches and practical guidelines.* New York: Longman.

APPENDICES

APPENDIX A

UEW Intern Teaching Evaluation Form

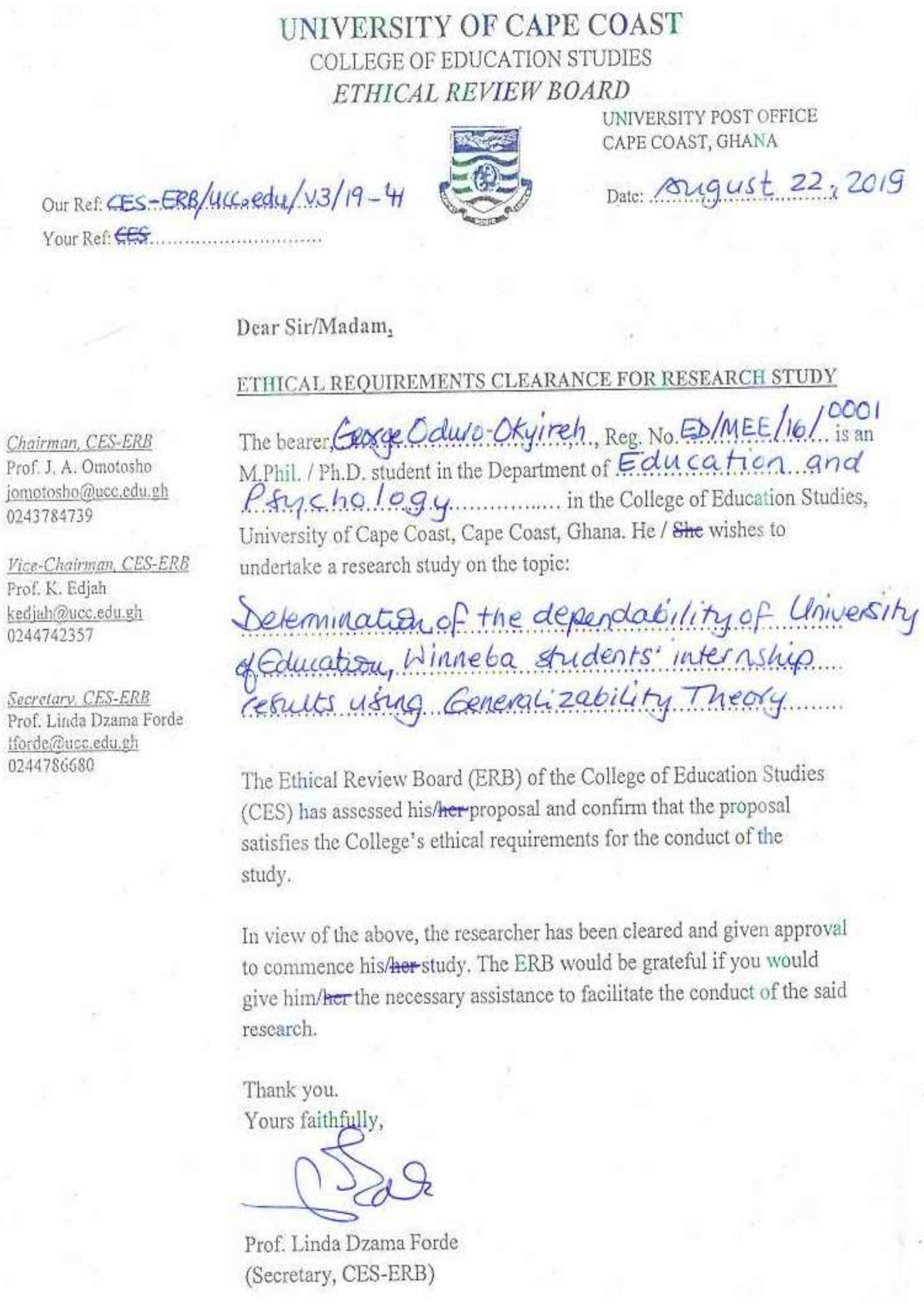
		SCORES				
PLANNING AND PREPARATION		**0**	**1**	**2**	**3**	**4**
1.	Exhibits knowledge of subject matter					
2.	Objectives are "SMART" and align instructional strategies with lesson objectives					
3.	Content connects with and challenges students' present knowledge, skills and values					
INSTRUCTIONAL SKILLS						
1.	States purpose, objectives, and procedures for lessons.					
2.	Gives procedural and instructional directions clearly.					
3.	Uses a range of strategies for whole class, small group and individual teaching/learning.					
4.	Motivates students.					
5.	Relates lesson to prior knowledge and life experience.					
6.	Presents lesson in a systematic manner.					
7.	Uses effective questioning techniques of the level of students.					
8.	Engages students in critical thinking and problem solving.					
9.	Uses techniques that modify and extend student learning.					
10.	Engages students in lesson closure.					
CLASSROOM MANAGEMENT						
1.	Manages classroom routines effectively.					
2.	Respects diversity among students.					
3.	Maintains Positive Rapport with students.					
	Knows each student as an individual.					
COMMUNICATION SKILLS						
1.	Communicates with confidence and enthusiasm.					
2.	Communicates at students' level of understanding.					
3.	Uses accurate non – verbal, oral/sign and written communication					
4.	Projects voice/hand shapes/orientation appropriately.					
EVALUATION						
1.	Monitors student's participation and progress.					
2.	Provides immediate and constructive feedback.					
3.	Bases evaluation on instructional goals/objectives.					
4.	Uses formal/informal assessment strategies to assess student learning before/during/after instruction to enhance learning.					

Total

Score………………Grade………………Signature………………………

APPENDIX B

Letter of Ethical Review



UNIVERSITY OF CAPE COAST
COLLEGE OF EDUCATION STUDIES
ETHICAL REVIEW BOARD

UNIVERSITY POST OFFICE
CAPE COAST, GHANA

Our Ref: CES-ERB/ucc.edu/v3/19-41

Your Ref: CES

Date: August 22, 2019

Chairman, CES-ERB
Prof. J. A. Omotosho
jomotosho@ucc.edu.gh
0243784739

Vice-Chairman, CES-ERB
Prof. K. Edjah
kedjah@ucc.edu.gh
0244742357

Secretary, CES-ERB
Prof. Linda Dzama Forde
lforde@ucc.edu.gh
0244786680

Dear Sir/Madam,

ETHICAL REQUIREMENTS CLEARANCE FOR RESEARCH STUDY

The bearer, George Oduro-Okyireh, Reg. No. ED/MEE/16/0001 is an M.Phil. / Ph.D. student in the Department of 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:

Determination of the dependability of University of Education, Winneba students' internship results using Generalizability Theory

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.

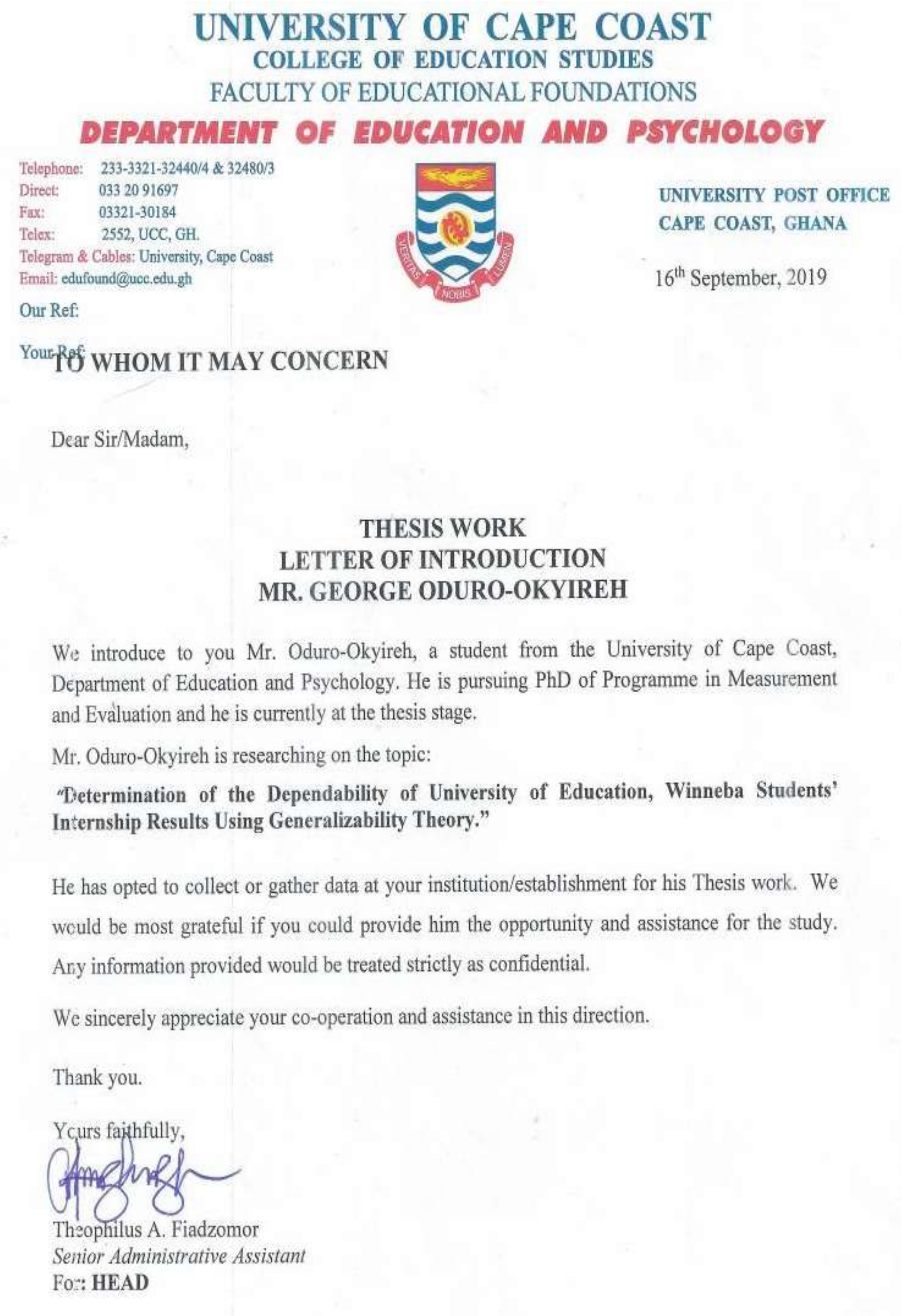
In view of the above, the researcher has been cleared and given approval to commence his/~~her~~ study. The ERB would be grateful if you would give him/~~her~~ the necessary assistance to facilitate the conduct of the said research.

Thank you.
Yours faithfully,

Prof. Linda Dzama Forde
(Secretary, CES-ERB)

APPENDIX C

Letter of Introduction



UNIVERSITY OF CAPE COAST
COLLEGE OF EDUCATION STUDIES
FACULTY OF EDUCATIONAL FOUNDATIONS
DEPARTMENT OF EDUCATION AND PSYCHOLOGY

Telephone: 233-3321-32440/4 & 32480/3
Direct: 033 20 91697
Fax: 03321-30184
Telex: 2552, UCC, GH.
Telegram & Cables: University, Cape Coast
Email: edufound@ucc.edu.gh

UNIVERSITY POST OFFICE
CAPE COAST, GHANA

16th September, 2019

Our Ref:

Your Ref:

TO WHOM IT MAY CONCERN

Dear Sir/Madam,

THESIS WORK
LETTER OF INTRODUCTION
MR. GEORGE ODURO-OKYIREH

We introduce to you Mr. Oduro-Okyireh, a student from the University of Cape Coast, Department of Education and Psychology. He is pursuing PhD of Programme in Measurement and Evaluation and he is currently at the thesis stage.

Mr. Oduro-Okyireh is researching on the topic:

"Determination of the Dependability of University of Education, Winneba Students' Internship Results Using Generalizability Theory."

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,

Theophilus A. Fiadzomor
Senior Administrative Assistant
For: **HEAD**

APPENDIX E_1

G and D Study Analyses of Mentors' Results for 2015/2016 Academic Year

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-03-01 11:45]

INTERNSHIP DATA FROM MENTORS - APPLIED SCIENCE, 2015/2016. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	247	INF	
OCCASIONS	O	3	INF	

Analysis of variance

Source	SS	Df	MS	Components				
				Random	Mixed	Corrected	%	SE
P	27046.6181	246	109.9456	27.1878	27.1878	27.1878	48.8	3.3457
O	123.9298	2	61.9649	0.1360	0.1360	0.1360	0.2	0.1775
PO	13964.0702	492	28.3823	28.3823	28.3823	28.3823	51.0	1.8059
Total	41134.6181	740					100%	

G Study Table (Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% Relative	Absolute error variance	% absolute
P	27.1878		.....		.....	
	.....	O	.....		0.0453	0.5
	.....	PO	9.4608	100.0	9.4608	99.5
Sum of variances	27.1878		9.4608	100%	9.5061	100%
Standard Deviation	5.2142		Relative SE: 3.0758		Absolute SE: 3.0832	
Coef_G relative	0.74					
Coef_G absolute	0.74					

Grand mean for levels used: 81.6586
Variance error of the mean for levels used: 0.1937
Standard error of the grand mean: 0.4401

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9825

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	247	INF	247	INF	247	INF	247	INF	247	INF	247	INF
O	3	INF	2	INF	3	INF	4	INF	5	INF	6	INF
Observ.		741		494		741		988		1235		1482
Coef_G rel.		0.7419		0.6570		0.7419		0.7930		0.8273		0.8518
Rounded		0.74		0.66		0.74		0.79		0.83		0.85
Coef_G abs.		0.7409		0.6560		0.7409		0.7922		0.8266		0.8512
Rounded		0.74		0.66		0.74		0.79		0.83		0.85
Rel. Err. Var.		9.4608		14.1911		9.4608		7.0956		5.6765		4.7304
Rel. Std. Err. of M.		3.0758		3.7671		3.0758		2.6638		2.3825		2.1749
Abs. Err. Var.		9.5061		14.2591		9.5061		7.1296		5.7036		4.7530
Abs. Std. Err. of M.		3.0832		3.7761		3.0832		2.6701		2.3882		2.1801

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen-
[2020-04-25 14:55]

INTERNSHIP DATA FROM MENTORS - BUSINESS EDUCATION, 2015/2016 ACADEMIC YEAR.
DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	800	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	48858.4296	799	61.1495	15.6363	15.6363	15.6363	44.2	1.0323
O	8889.8008	2	4444.9004	5.5383	5.5383	5.5383	15.6	3.9288
PO	22756.1992	1598	14.2404	14.2404	14.2404	14.2404	40.2	0.5035
Total	80504.4296	2399					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	15.6363		.....		.....	
	.....	O	.....		1.8461	28.0
	.....	PO	4.7468	100.0	4.7468	72.0
Sum of variances	15.6363		4.7468	100%	6.5929	100%
Standard deviation	3.9543		Relative SE: 2.1787		Absolute SE: 2.5677	
Coef_G relative	0.77					
Coef_G absolute	0.70					

Grand mean for levels used: 79.9221
Variance error of the mean for levels used: 1.8716
Standard error of the grand mean: 1.3681

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9928

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	800	INF	800	INF	800	INF	800	INF	800	INF	800	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.	2400		800		1600		3200		4000		4800	
Coef_G rel.	0.7671		0.5234		0.6871		0.8145		0.8459		0.8682	
rounded	0.77		0.52		0.69		0.81		0.85		0.87	
Coef_G abs.	0.7034		0.4415		0.6126		0.7597		0.7981		0.8259	
rounded	0.70		0.44		0.61		0.76		0.80		0.83	
Rel. Err. Var.	4.7468		14.2404		7.1202		3.5601		2.8481		2.3734	
Rel. Std. Err. of M.	2.1787		3.7736		2.6684		1.8868		1.6876		1.5406	
Abs. Err. Var.	6.5929		19.7788		9.8894		4.9447		3.9558		3.2965	
Abs. Std. Err. of M.	2.5677		4.4473		3.1447		2.2237		1.9889		1.8156	

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-04-21 10:41]

INTERNSHIP DATA FROM MENTORS - ENGLISH AND COMMUNICATION, 2015/2016 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	110	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	Df	MS	Random	Mixed	Corrected	%	SE
P	7503.4212	109	68.8387	17.6331	17.6331	17.6331	36.8	3.1215
O	3187.8970	2	1593.9485	14.3455	14.3455	14.3455	29.9	10.2463
PO	3474.7697	218	15.9393	15.9393	15.9393	15.9393	33.3	1.5198
Total	14166.0879	329					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	17.6331		.....		.....	
	.....	O	.....		4.7818	47.4
	.....	PO	5.3131	100.0	5.3131	52.6
Sum of variances	17.6331		5.3131	100%	10.0949	100%
Standard deviation	4.1992		Relative SE: 2.3050		Absolute SE: 3.1773	
Coef_G relative	0.77					
Coef_G absolute	0.64					

Grand mean for levels used: 78.6394
Variance error of the mean for levels used: 4.9904
Standard error of the grand mean: 2.2339

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9880

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	110	INF	110	INF	110	INF	110	INF	110	INF	110	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.	330		110		220		440		550		660	
Coef_G rel.	0.7685		0.5252		0.6887		0.8157		0.8469		0.8691	
Rounded	0.77		0.53		0.69		0.82		0.85		0.87	
Coef_G abs.	0.6359		0.3680		0.5380		0.6996		0.7443		0.7775	
Rounded	0.64		0.37		0.54		0.70		0.74		0.78	
Rel. Err. Var.	5.3131		15.9393		7.9697		3.9848		3.1879		2.6566	
Rel. Std. Err. of M.	2.3050		3.9924		2.8231		1.9962		1.7855		1.6299	
Abs. Err. Var.	10.0949		30.2848		15.1424		7.5712		6.0570		5.0475	
Abs. Std. Err. of M.	3.1773		5.5032		3.8913		2.7516		2.4611		2.2467	

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-03-19 07:51]

INTERNSHIP DATA FROM MENTORS – FOREIGN LANGUAGES AND LINGUISTICS, 2015/2016 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	113	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	13945.6283	112	124.5145	31.0233	31.0233	31.0233	49.5	5.5852
O	97.7050	2	48.8525	0.1541	0.1541	0.1541	0.2	0.3068
PO	7043.6283	224	31.4448	31.4448	31.4448	31.4448	50.2	2.9581
Total	21086.9617	338					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% Absolute
P	31.0233		.....		.....	
	.....	O	.....		0.0514	0.5
	.....	PO	10.4816	100.0	10.4816	99.5
Sum of variances	31.0233		10.4816	100%	10.5329	100%
Standard deviation	5.5699		Relative SE: 3.2375		Absolute SE: 3.2454	
Coef_G relative	0.75					
Coef_G absolute	0.75					

Grand mean for levels used: 81.7758
Variance error of the mean for levels used: 0.4186
Standard error of the grand mean: 0.6470

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9900

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	113	INF	113	INF	113	INF	113	INF	113	INF	113	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.	339		113		226		452		565		678	
Coef_G rel.	0.7475		0.4966		0.6637		0.7978		0.8315		0.8555	
rounded	0.75		0.50		0.66		0.80		0.83		0.86	
Coef_G abs.	0.7465		0.4954		0.6626		0.7970		0.8308		0.8549	
rounded	0.75		0.50		0.66		0.80		0.83		0.85	
Rel. Err. Var.	10.4816		31.4448		15.7224		7.8612		6.2890		5.2408	
Rel. Std. Err. of M.	3.2375		5.6076		3.9651		2.8038		2.5078		2.2893	
Abs. Err. Var.	10.5329		31.5988		15.7994		7.8997		6.3198		5.2665	
Abs. Std. Err. of M.	3.2454		5.6213		3.9748		2.8106		2.5139		2.2949	

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-04-26 07:35]

INTERNSHIP DATA FROM MENTORS - NATURAL SCIENCE, 2015/2016 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	200	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	22055.8400	199	110.8334	28.9441	28.9441	28.9441	54.6	3.7284
O	74.9233	2	37.4617	0.0673	0.0673	0.0673	0.1	0.1327
PO	9552.4100	398	24.0010	24.0010	24.0010	24.0010	45.3	1.6971
Total	31683.1733	599					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	28.9441		.....		.....	
	.....	O	.....		0.0224	0.3
	.....	PO	8.0003	100.0	8.0003	99.7
Sum of variances	28.9441		8.0003	100%	8.0228	100%
Standard deviation	5.3800		Relative SE: 2.8285		Absolute SE: 2.8325	
Coef_G relative	0.78					
Coef_G absolute	0.78					

Grand mean for levels used: 80.7733
Variance error of the mean for levels used: 0.2072
Standard error of the grand mean: 0.4551

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9918

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	200	INF	200	INF	200	INF	200	INF	200	INF	200	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		600		200		400		800		1000		1200
Coef_G rel.		0.7834		0.5467		0.7069		0.8283		0.8577		0.8786
rounded		0.78		0.55		0.71		0.83		0.86		0.88
Coef_G abs.		0.7830		0.5460		0.7063		0.8279		0.8574		0.8783
rounded		0.78		0.55		0.71		0.83		0.86		0.88
Rel. Err. Var.		8.0003		24.0010		12.0005		6.0003		4.8002		4.0002
Rel. Std. Err. of M.		2.8285		4.8991		3.4642		2.4495		2.1909		2.0000
Abs. Err. Var.		8.0228		24.0683		12.0342		6.0171		4.8137		4.0114
Abs. Std. Err. of M.		2.8325		4.9059		3.4690		2.4530		2.1940		2.0028

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-04-10 10:43]

INTERNSHIP DATA FROM MENTORS - SOCIAL SCIENCE, 2015/2016 ACADEMIC YEAR.
DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	577	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	Df	MS	Random	Mixed	Corrected	%	SE
P	47873.1600	576	83.1131	18.6144	18.6144	18.6144	40.0	1.6730
O	762.5176	2	381.2588	0.6135	0.6135	0.6135	1.3	0.4672
PO	31414.8157	1152	27.2698	27.2698	27.2698	27.2698	58.6	1.1353
Total	80050.4934	1730					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	18.6144		.....		.....	
	.....	O	.....		0.2045	2.2
	.....	PO	9.0899	100.0	9.0899	97.8
Sum of variances	18.6144		9.0899	100%	9.2944	100%
Standard deviation	4.3144		Relative SE: 3.0150		Absolute SE: 3.0487	
Coef_G relative	0.67					
Coef_G absolute	0.67					

Grand mean for levels used: 84.4639
Variance error of the mean for levels used: 0.2525
Standard error of the grand mean: 0.5025

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9924

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	577	INF	577	INF	577	INF	577	INF	577	INF	577	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.	1731		577		1154		2308		2885		3462	
Coef_G rel.	0.6719		0.4057		0.5772		0.7319		0.7734		0.8038	
rounded	0.67		0.41		0.58		0.73		0.77		0.80	
Coef_G abs.	0.6670		0.4003		0.5718		0.7275		0.7695		0.8002	
rounded	0.67		0.40		0.57		0.73		0.77		0.80	
Rel. Err. Var.	9.0899		27.2698		13.6349		6.8175		5.4540		4.5450	
Rel. Std. Err. of M.	3.0150		5.2220		3.6925		2.6110		2.3354		2.1319	
Abs. Err. Var.	9.2944		27.8833		13.9417		6.9708		5.5767		4.6472	
Abs. Std. Err. of M.	3.0487		5.2805		3.7339		2.6402		2.3615		2.1557	

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-04-21 07:50]

INTERNSHIP DATA FROM MENTORS - TECHNICAL EDUCATION, 2015/2016 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	470	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	26012.7972	469	55.4644	12.7676	12.7676	12.7676	32.6	1.2333
O	8660.5291	2	4330.2645	9.1768	9.1768	9.1768	23.5	6.5148
PO	16097.4709	938	17.1615	17.1615	17.1615	17.1615	43.9	0.7916
Total	50770.7972	1409					100%	

G Study Table (Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% Absolute
P	12.7676		.....		.....	
	.....	O	.....		3.0589	34.8
	.....	PO	5.7205	100.0	5.7205	65.2
Sum of variances	12.7676		5.7205	100%	8.7794	100%
Standard deviation	3.5732		Relative SE: 2.3918		Absolute SE: 2.9630	
Coef_G relative	0.69					
Coef_G absolute	0.59					

Grand mean for levels used: 80.2652
Variance error of the mean for levels used: 3.0983
Standard error of the grand mean: 1.7602

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9906

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	470	INF	470	INF	470	INF	470	INF	470	INF	470	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		1410		470		940		1880		2350		2820
Coef_G rel.		0.6906		0.4266		0.5981		0.7485		0.7881		0.8170
rounded		0.69		0.43		0.60		0.75		0.79		0.82
Coef_G abs.		0.5925		0.3265		0.4923		0.6598		0.7079		0.7441
rounded		0.59		0.33		0.49		0.66		0.71		0.74
Rel. Err. Var.		5.7205		17.1615		8.5807		4.2904		3.4323		2.8602
Rel. Std. Err. of M.		2.3918		4.1426		2.9293		2.0713		1.8526		1.6912
Abs. Err. Var.		8.7794		26.3383		13.1691		6.5846		5.2677		4.3897
Abs. Std. Err. of M.		2.9630		5.1321		3.6289		2.5660		2.2951		2.0952

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-04-20 19:23]

INTERNSHIP DATA FROM MENTORS - VOCATIONAL EDUCATION, 2015/2016 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	79	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	10213.2911	78	130.9396	32.7560	32.7560	32.7560	46.0	7.0091
O	983.9072	2	491.9536	5.8137	5.8137	5.8137	8.2	4.4036
PO	5096.7595	156	32.6715	32.6715	32.6715	32.6715	45.9	3.6758
Total	16293.9578	236					100%	

G Study Table (Measurement design P/O)

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% Absolute
P	32.7560		.....		.....	
	.....	O	.....		1.9379	15.1
	.....	PO	10.8905	100.0	10.8905	84.9
Sum of variances	32.7560		10.8905	100%	12.8284	100%
Standard deviation	5.7233		Relative SE: 3.3001		Absolute SE: 3.5817	
Coef_G relative	0.75					
Coef_G absolute	0.72					

Grand mean for levels used: 78.9072
Variance error of the mean for levels used: 2.4904
Standard error of the grand mean: 1.5781

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9854

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	79	INF	79	INF	79	INF	79	INF	79	INF	79	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		237		79		158		316		395		474
Coef_G rel.		0.7505		0.5006		0.6672		0.8004		0.8337		0.8575
rounded		0.75		0.50		0.67		0.80		0.83		0.86
Coef_G abs.		0.7186		0.4598		0.6299		0.7730		0.8097		0.8362
rounded		0.72		0.46		0.63		0.77		0.81		0.84
Rel. Err. Var.		10.8905		32.6715		16.3358		8.1679		6.5343		5.4453
Rel. Std. Err. of M.		3.3001		5.7159		4.0418		2.8580		2.5562		2.3335
Abs. Err. Var.		12.8284		38.4852		19.2426		9.6213		7.6970		6.4142
Abs. Std. Err. of M.		3.5817		6.2036		4.3866		3.1018		2.7744		2.5326

APPENDIX E_2

G and D Study Analyses of Mentors' Results for 2016/2017 Academic Year

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-03-06 10:57]

INTERNSHIP DATA FROM MENTORS - APPLIED SCIENCE, 2016/2017 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	309	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	Df	MS	Random	Mixed	Corrected	%	SE
P	46005.5793	308	149.3688	32.9566	32.9566	32.9566	39.2	4.1122
O	424.0065	2	212.0032	0.5227	0.5227	0.5227	0.6	0.4852
PO	31107.3269	616	50.4989	50.4989	50.4989	50.4989	60.1	2.8728
Total	77536.9126	926					100%	

G Study Table (Measurement design P/O)

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	32.9566		.....		.....	
	.....	O	.....		0.1742	1.0
	.....	PO	16.8330	100.0	16.8330	99.0
Sum of variances	32.9566		16.8330	100%	17.0072	100%
Standard deviation	5.7408		Relative SE: 4.1028		Absolute SE: 4.1240	
Coef_G relative	0.66					
Coef_G absolute	0.66					

Grand mean for levels used: 80.3236
Variance error of the mean for levels used: 0.3354
Standard error of the grand mean: 0.5791

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9827

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	309	INF	309	INF	309	INF	309	INF	309	INF	309	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		927		309		618		1236		1545		1854
Coef_G rel.		0.6619		0.3949		0.5662		0.7230		0.7654		0.7966
rounded		0.66		0.39		0.57		0.72		0.77		0.80
Coef_G abs.		0.6596		0.3924		0.5637		0.7210		0.7636		0.7949
rounded		0.66		0.39		0.56		0.72		0.76		0.79
Rel. Err. Var.		16.8330		50.4989		25.2495		12.6247		10.0998		8.4165
Rel. Std. Err. of M.		4.1028		7.1063		5.0249		3.5531		3.1780		2.9011
Abs. Err. Var.		17.0072		51.0216		25.5108		12.7554		10.2043		8.5036
Abs. Std. Err. of M.		4.1240		7.1429		5.0508		3.5715		3.1944		2.9161

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-04-15 14:00]

INTERNSHIP DATA FROM MENTORS – BUSINESS EDUCATION, 2016/2017 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	633	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	48906.0042	632	77.3829	20.8848	20.8848	20.8848	53.5	1.4618
O	4337.8210	2	2168.9105	3.4031	3.4031	3.4031	8.7	2.4228
PO	18616.8457	1264	14.7285	14.7285	14.7285	14.7285	37.7	0.5854
Total	71860.6709	1898					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	20.8848		.....		.....	
	.....	O	.....		1.1344	18.8
	.....	PO	4.9095	100.0	4.9095	81.2
Sum of variances	20.8848		4.9095	100%	6.0439	100%
Standard deviation	4.5700		Relative SE: 2.2157		Absolute SE: 2.4584	
Coef_G relative	0.81					
Coef_G absolute	0.78					

Grand mean for levels used: 80.1243
Variance error of the mean for levels used: 1.1751
Standard error of the grand mean: 1.0840

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9935

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	633	INF	633	INF	633	INF	633	INF	633	INF	633	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.	1899		633		1266		2532		3165		3798	
Coef_G rel.	0.8097		0.5864		0.7393		0.8501		0.8764		0.8948	
rounded	0.81		0.59		0.74		0.85		0.88		0.89	
Coef_G abs.	0.7756		0.5353		0.6973		0.8217		0.8521		0.8736	
rounded	0.78		0.54		0.70		0.82		0.85		0.87	
Rel. Err. Var.	4.9095		14.7285		7.3643		3.6821		2.9457		2.4548	
Rel. Std. Err. of M.	2.2157		3.8378		2.7137		1.9189		1.7163		1.5668	
Abs. Err. Var.	6.0439		18.1316		9.0658		4.5329		3.6263		3.0219	
Abs. Std. Err. of M.	2.4584		4.2581		3.0110		2.1291		1.9043		1.7384	

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-04-16 09:41

INTERNSHIP DATA FROM MENTORS - ENGLISH AND COMMUNICATION EDUCATION, 2016/2017 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	61	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	4881.7705	60	81.3628	19.6762	19.6762	19.6762	44.7	4.9635
O	286.5683	2	143.2842	1.9828	1.9828	1.9828	4.5	1.6616
PO	2680.0984	120	22.3342	22.3342	22.3342	22.3342	50.8	2.8596
Total	7848.4372	182					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	19.6762		.....		.....	
	.....	O	.....		0.6609	8.2
	.....	PO	7.4447	100.0	7.4447	91.8
Sum of variances	19.6762		7.4447	100%	8.1056	100%
Standard deviation	4.4358		Relative SE: 2.7285		Absolute SE: 2.8470	
Coef_G relative	0.73					
Coef_G absolute	0.71					

Grand mean for levels used: 79.7486
Variance error of the mean for levels used: 1.1055
Standard error of the grand mean: 1.0514

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9911

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	61	INF	490	INF	490	INF	490	INF	490	INF	490	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		183		490		980		1960		2450		2940
Coef_G rel.		0.7255		0.4684		0.6379		0.7790		0.8150		0.8409
rounded		0.73		0.47		0.64		0.78		0.81		0.84
Coef_G abs.		0.7082		0.4473		0.6181		0.7640		0.8018		0.8292
rounded		0.71		0.45		0.62		0.76		0.80		0.83
Rel. Err. Var.		7.4447		22.3342		11.1671		5.5835		4.4668		3.7224
Rel. Std. Err. of M.		2.7285		4.7259		3.3417		2.3630		2.1135		1.9293
Abs. Err. Var.		8.1056		24.3169		12.1585		6.0792		4.8634		4.0528
Abs. Std. Err. of M.		2.8470		4.9312		3.4869		2.4656		2.2053		2.0132

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-03-19 11:03]

INTERNSHIP DATA FROM MENTORS – FOREIGN LANGUAGES AN D LINGUISTICS, 2016/2017 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	342	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	43189.1930	341	126.6545	34.1457	34.1457	34.1457	58.3	3.2532
O	190.3392	2	95.1696	0.2075	0.2075	0.2075	0.4	0.1968
PO	16516.3275	682	24.2175	24.2175	24.2175	24.2175	41.3	1.3095
Total	59895.8596	1025					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	34.1457		.....		.....	
	.....	O	.....		0.0692	0.8
	.....	PO	8.0725	100.0	8.0725	99.2
Sum of variances	34.1457		8.0725	100%	8.1417	100%
Standard deviation	5.8434		Relative SE: 2.8412		Absolute SE: 2.8534	
Coef_G relative	0.81					
Coef_G absolute	0.81					

Grand mean for levels used: 81.3216
Variance error of the mean for levels used: 0.1926
Standard error of the grand mean: 0.4389

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9920

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	342	INF	342	INF	342	INF	342	INF	342	INF	342	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.	1026		342		684		1368		1710		2052	
Coef_G rel.	0.8088		0.5851		0.7382		0.8494		0.8758		0.8943	
rounded	0.81		0.59		0.74		0.85		0.88		0.89	
Coef_G abs.	0.8075		0.5830		0.7366		0.8483		0.8748		0.8935	
rounded	0.81		0.58		0.74		0.85		0.87		0.89	
Rel. Err. Var.	8.0725		24.2175		12.1087		6.0544		4.8435		4.0362	
Rel. Std. Err. of M.	2.8412		4.9211		3.4798		2.4606		2.2008		2.0090	
Abs. Err. Var.	8.1417		24.4250		12.2125		6.1062		4.8850		4.0708	
Abs. Std. Err. of M.	2.8534		4.9422		3.4946		2.4711		2.2102		2.0176	

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-03-06 04:31]

INTERNSHIP DATA FROM MENTORS – NATURAL SCIENCE, 2016/2017 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	276	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	35126.2464	275	127.7318	34.7923	34.7923	34.7923	58.9	3.6481
O	549.4855	2	274.7428	0.9108	0.9108	0.9108	1.5	0.7039
PO	12845.1812	550	23.3549	23.3549	23.3549	23.3549	39.5	1.4058
Total	48520.9130	827					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	34.7923		.....		.....	
	.....	O	.....		0.3036	3.8
	.....	PO	7.7850	100.0	7.7850	96.2
Sum of variances	34.7923		7.7850	100%	8.0886	100%
Standard deviation	5.8985		Relative SE: 2.7902		Absolute SE: 2.8440	
Coef_G relative	0.82					
Coef_G absolute	0.81					

Grand mean for levels used: 80.7029
Variance error of the mean for levels used: 0.4579
Standard error of the grand mean: 0.6767

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9918

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	276	INF	276	INF	276	INF	276	INF	276	INF	276	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.	828		276		552		1104		1380		1656	
Coef_G rel.	0.8172		0.5983		0.7487		0.8563		0.8816		0.8994	
rounded	0.82		0.60		0.75		0.86		0.88		0.90	
Coef_G abs.	0.8114		0.5891		0.7414		0.8515		0.8776		0.8959	
rounded	0.81		0.59		0.74		0.85		0.88		0.90	
Rel. Err. Var.	7.7850		23.3549		11.6774		5.8387		4.6710		3.8925	
Rel. Std. Err. of M.	2.7902		4.8327		3.4172		2.4163		2.1612		1.9729	
Abs. Err. Var.	8.0886		24.2657		12.1329		6.0664		4.8531		4.0443	
Abs. Std. Err. of M.	2.8440		4.9260		3.4832		2.4630		2.2030		2.0110	

File C:\Program Files (x86)\EduG - 6.1e\Data\ON-CAMPUS TEACHING PRACTICE.gen - [2020-04-27 15:34]
INTERNSHIP DATA FROM MENTORS - SOCIAL SCIENCE, 2016/2017 ACADEMIC YEAR.
DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	835	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	70628.0080	834	84.6859	19.3517	19.3517	19.3517	41.8	1.4145
O	501.9457	2	250.9729	0.2687	0.2687	0.2687	0.6	0.2125
PO	44420.0543	1668	26.6307	26.6307	26.6307	26.6307	57.6	0.9216
Total	115550.0080	2504					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	19.3517		.....		.....	
	.....	O	.....		0.0896	1.0
	.....	PO	8.8769	100.0	8.8769	99.0
Sum of Variances	19.3517		8.8769	100%	8.9665	100%
Standard Deviation	4.3991		Relative SE: 2.9794		Absolute SE: 2.9944	
Coef_G relative	0.69					
Coef_G absolute	0.68					

Grand mean for levels used: 84.6267
Variance error of the mean for levels used: 0.1234
Standard error of the grand mean: 0.3512

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9927

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	835	INF	835	INF	835	INF	835	INF	835	INF	835	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.	2505		835		1670		3340		4175		5010	
Coef_G rel.	0.6855		0.4209		0.5924		0.7440		0.7842		0.8134	
Rounded	0.69		0.42		0.59		0.74		0.78		0.81	
Coef_G abs.	0.6834		0.4184		0.5900		0.7421		0.7825		0.8119	
rounded	0.68		0.42		0.59		0.74		0.78		0.81	
Rel. Err. Var.	8.8769		26.6307		13.3154		6.6577		5.3261		4.4385	
Rel. Std. Err. of M.	2.9794		5.1605		3.6490		2.5802		2.3078		2.1068	
Abs. Err. Var.	8.9665		26.8994		13.4497		6.7249		5.3799		4.4832	
Abs. Std. Err. of M.	2.9944		5.1865		3.6674		2.5932		2.3195		2.1174	

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-04-16 06:18]

INTERNSHIP DATA FROM MENTORS – TECHNICAL EDUCATION, 2016/2017 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	490	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	48209.2000	489	98.5873	25.0972	25.0972	25.0972	46.6	2.1265
O	5395.4136	2	2697.7068	5.4580	5.4580	5.4580	10.1	3.8930
PO	22783.2531	978	23.2958	23.2958	23.2958	23.2958	43.3	1.0524
Total	76387.8667	1469					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% Absolute
P	25.0972		.....		.....	
	.....	O	.....		1.8193	19.0
	.....	PO	7.7653	100.0	7.7653	81.0
Sum of variances	25.0972		7.7653	100%	9.5846	100%
Standard deviation	5.0097		Relative SE: 2.7866		Absolute SE: 3.0959	
Coef_G relative	0.76					
Coef_G absolute	0.72					

Grand mean for levels used: 79.9905
Variance error of the mean for levels used: 1.8864
Standard error of the grand mean: 1.3735

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9897

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	490	INF	490	INF	490	INF	490	INF	490	INF	490	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.	1470		490		980		1960		2450			2940
Coef_G rel.	0.7637		0.5186		0.6830		0.8117		0.8434		0.8660	
rounded	0.76		0.52		0.68		0.81		0.84			0.87
Coef_G abs.	0.7236		0.4660		0.6358		0.7773		0.8136		0.8397	
rounded	0.72		0.47		0.64		0.78		0.81			0.84
Rel. Err. Var.	7.7653		23.2958		11.6479		5.8239		4.6592		3.8826	
Rel. Std. Err. of M.	2.7866		4.8266		3.4129		2.4133		2.1585		1.9704	
Abs. Err. Var.	9.5846		28.7537		14.3769		7.1884		5.7507		4.7923	
Abs. Std. Err. of M.	3.0959		5.3623		3.7917		2.6811		2.3981		2.1891	

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-04-10 16:45]

INTERNSHIP DATA FROM MENTORS - VOCATIONAL, 2016/2017 ACADEMIC YEAR.
DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	109	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	9129.8593	108	84.5357	23.6924	23.6924	23.6924	58.9	3.8238
O	690.9786	2	345.4893	3.0462	3.0462	3.0462	7.6	2.2413
PO	2907.0214	216	13.4584	13.4584	13.4584	13.4584	33.5	1.2891
Total	12727.8593	326					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% Absolute
P	23.6924		.....		.....	
	.....	O	.....		1.0154	18.5
	.....	PO	4.4861	100.0	4.4861	81.5
Sum of variances	23.6924		4.4861	100%	5.5015	100%
Standard deviation	4.8675		Relative SE: 2.1181		Absolute SE: 2.3455	
Coef_G relative	0.84					
Coef_G absolute	0.81					

Grand mean for levels used: 80.7920
Variance error of the mean for levels used: 1.2739
Standard error of the grand mean: 1.1287

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9944

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	109	INF	109	INF	109	INF	109	INF	109	INF	109	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.	327		109		218		436		545		654	
Coef_G rel.	0.8408		0.6377		0.7788		0.8756		0.8980		0.9135	
rounded	0.84		0.64		0.78		0.88		0.90		0.91	
Coef_G abs.	0.8116		0.5894		0.7417		0.8517		0.8777		0.8960	
rounded	0.81		0.59		0.74		0.85		0.88		0.90	
Rel. Err. Var.	4.4861		13.4584		6.7292		3.3646		2.6917		2.2431	
Rel. Std. Err. of M.	2.1181		3.6686		2.5941		1.8343		1.6406		1.4977	
Abs. Err. Var.	5.5015		16.5046		8.2523		4.1261		3.3009		2.7508	
Abs. Std. Err. of M.	2.3455		4.0626		2.8727		2.0313		1.8168		1.6585	

APPENDIX E_3

G and D Study Analyses of Mentors' Results for 2017/2018 Academic Year

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-03-07 11:24]

INTERNSHIP DATA FROM MENTORS - APPLIED SCIENCE, 2017/2018 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	342	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	Df	MS	Random	Mixed	Corrected	%	SE
P	36792.3353	341	107.8954	25.8224	25.8224	25.8224	45.0	2.8005
O	807.9201	2	403.9600	1.0922	1.0922	1.0922	1.9	0.8352
PO	20752.0799	682	30.4283	30.4283	30.4283	30.4283	53.1	1.6454
Total	58352.3353	1025					100%	

G Study Table (Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	25.8224		.....		.....	
	.....	O	.....		0.3641	3.5
	.....	PO	10.1428	100.0	10.1428	96.5
Sum of variances	25.8224		10.1428	100%	10.5068	100%
Standard deviation	5.0816		Relative SE:	3.1848	Absolute SE:	3.2414
Coef_G relative	0.72					
Coef_G absolute	0.71					

Grand mean for levels used: 82.3021
Variance error of the mean for levels used: 0.4692
Standard error of the grand mean: 0.6850

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9903

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	342	INF	342	INF	342	INF	342	INF	342	INF	342	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		1026		342		684		1368		1710		2052
Coef_G rel.		0.7180		0.4591		0.6293		0.7724		0.8093		0.8358
rounded		0.72		0.46		0.63		0.77		0.81		0.84
Coef_G abs.		0.7108		0.4503		0.6210		0.7662		0.8038		0.8309
rounded		0.71		0.45		0.62		0.77		0.80		0.83
Rel. Err. Var.		10.1428		30.4283		15.2141		7.6071		6.0857		5.0714
Rel. Std. Err. of M.		3.1848		5.5162		3.9005		2.7581		2.4669		2.2520
Abs. Err. Var.		10.5068		31.5205		15.7602		7.8801		6.3041		5.2534
Abs. Std. Err. of M.		3.2414		5.6143		3.9699		2.8072		2.5108		2.2920

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-04-18 09:11

INTERNSHIP DATA FROM MENTORS - BUSINESS EDUCATION, 2017/2018 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	538	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	36701.4108	537	68.3453	16.7265	16.7265	16.7265	45.4	1.4121
O	2118.0260	2	1059.0130	1.9347	1.9347	1.9347	5.3	1.3919
PO	19509.9740	1074	18.1657	18.1657	18.1657	18.1657	49.3	0.7832
Total	58329.4108	1613					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% Relative	Absolute error variance	% Absolute
P	16.7265		.....		.....	
	.....	O	.....		0.6449	9.6
	.....	PO	6.0552	100.0	6.0552	90.4
Sum of variances	16.7265		6.0552	100%	6.7001	100%
Standard deviation	4.0898		Relative SE: 2.4607		Absolute SE: 2.5885	
Coef_G relative	0.73					
Coef_G absolute	0.71					

Grand mean for levels used: 74.5074
Variance error of the mean for levels used: 0.6872
Standard error of the grand mean: 0.8290

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9893

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	538	INF	538	INF	538	INF	538	INF	538	INF	538	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		1614		538		1076		2152		2690		3228
Coef_G rel.		0.7342		0.4794		0.6481		0.7865		0.8216		0.8467
rounded		0.73		0.48		0.65		0.79		0.82		0.85
Coef_G abs.		0.7140		0.4542		0.6247		0.7690		0.8062		0.8331
rounded		0.71		0.45		0.62		0.77		0.81		0.83
Rel. Err. Var.		6.0552		18.1657		9.0829		4.5414		3.6331		3.0276
Rel. Std. Err. of M.		2.4607		4.2621		3.0138		2.1311		1.9061		1.7400
Abs. Err. Var.		6.7001		20.1004		10.0502		5.0251		4.0201		3.3501
Abs. Std. Err. of M.		2.5885		4.4833		3.1702		2.2417		2.0050		1.8303

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-04-25 08:42]

INTERNSHIP DATA FROM MENTORS - ENGLISH AND COMMUNICATION, 2017/2018 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	150	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	8337.0311	149	55.9532	14.1380	14.1380	14.1380	47.6	2.1779
O	627.3244	2	313.6622	2.0008	2.0008	2.0008	6.7	1.4786
PO	4034.6756	298	13.5392	13.5392	13.5392	13.5392	45.6	1.1055
Total	12999.0311	449					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% Absolute
P	14.1380		.....		.....	
	.....	O	.....		0.6669	12.9
	.....	PO	4.5131	100.0	4.5131	87.1
Sum of variances	14.1380		4.5131	100%	5.1800	100%
Standard deviation	3.7601		Relative SE: 2.1244		Absolute SE: 2.2760	
Coef_G relative	0.76					
Coef_G absolute	0.73					

Grand mean for levels used: 75.8756
Variance error of the mean for levels used: 0.7913
Standard error of the grand mean: 0.8895

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9925

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	150	INF	150	INF	150	INF	150	INF	150	INF	150	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		450		150		300		600		750		900
Coef_G rel.		0.7580		0.5108		0.6762		0.8068		0.8393		0.8624
rounded		0.76		0.51		0.68		0.81		0.84		0.86
Coef_G abs.		0.7319		0.4764		0.6453		0.7844		0.8198		0.8452
rounded		0.73		0.48		0.65		0.78		0.82		0.85
Rel. Err. Var.		4.5131		13.5392		6.7696		3.3848		2.7078		2.2565
Rel. Std. Err. of M.		2.1244		3.6796		2.6018		1.8398		1.6456		1.5022
Abs. Err. Var.		5.1800		15.5400		7.7700		3.8850		3.1080		2.5900
Abs. Std. Err. of M.		2.2760		3.9421		2.7875		1.9710		1.7630		1.6093

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-03-18 20:44]

INTERNSHIP DATA FROM MENTORS – FOREIGN LANGUAGES AND LINGUISTICS, 2017/2018 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	353	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	43396.0548	352	123.2842	31.5788	31.5788	31.5788	52.4	3.1301
O	185.7167	2	92.8584	0.1822	0.1822	0.1822	0.3	0.1861
PO	20097.6166	704	28.5478	28.5478	28.5478	28.5478	47.3	1.5194
Total	63679.3881	1058					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% Absolute
P	31.5788		.....		.....	
	.....	O	.....		0.0607	0.6
	.....	PO	9.5159	100.0	9.5159	99.4
Sum of variances	31.5788		9.5159	100%	9.5766	100%
Standard deviation	5.6195		Relative SE: 3.0848		Absolute SE: 3.0946	
Coef_G relative	0.77					
Coef_G absolute	0.77					

Grand mean for levels used: 81.8385
Variance error of the mean for levels used: 0.1771
Standard error of the grand mean: 0.4209

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9909

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	353	INF	353	INF	353	INF	353	INF	353	INF	353	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		1059		353		706		1412		1765		2118
Coef_G rel.		0.7684		0.5252		0.6887		0.8157		0.8469		0.8691
rounded		0.77		0.53		0.69		0.82		0.85		0.87
Coef_G abs.		0.7673		0.5236		0.6873		0.8147		0.8461		0.8683
rounded		0.77		0.52		0.69		0.81		0.85		0.87
Rel. Err. Var.		9.5159		28.5478		14.2739		7.1369		5.7096		4.7580
Rel. Std. Err. of M.		3.0848		5.3430		3.7781		2.6715		2.3895		2.1813
Abs. Err. Var.		9.5766		28.7299		14.3650		7.1825		5.7460		4.7883
Abs. Std. Err. of M.		3.0946		5.3600		3.7901		2.6800		2.3971		2.1882

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-03-08 10:23]

INTERNSHIP DATA FROM MENTORS – NATURAL SCIENCE, 2017/2018 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	414	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	Df	MS	Random	Mixed	Corrected	%	SE
P	56100.1779	413	135.8358	36.6851	36.6851	36.6851	58.1	3.1715
O	590.5765	2	295.2882	0.6510	0.6510	0.6510	1.0	0.5044
PO	21294.7568	826	25.7806	25.7806	25.7806	25.7806	40.8	1.2670
Total	77985.5113	1241					100%	

G Study Table (Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% Absolute
P	36.6851		.....		.....	
	.....	O	.....		0.2170	2.5
	.....	PO	8.5935	100.0	8.5935	97.5
Sum of variances	36.6851		8.5935	100%	8.8105	100%
Standard deviation	6.0568		Relative SE: 2.9315		Absolute SE: 2.9683	
Coef_G relative	0.81					
Coef_G absolute	0.81					

Grand mean for levels used: 82.8172
Variance error of the mean for levels used: 0.3264
Standard error of the grand mean: 0.5713

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9921

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	414	INF	414	INF	414	INF	414	INF	414	INF	414	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		1242		414		828		1656		2070		2484
Coef_G rel.		0.8102		0.5873		0.7400		0.8506		0.8768		0.8952
rounded		0.81		0.59		0.74		0.85		0.88		0.90
Coef_G abs.		0.8063		0.5812		0.7352		0.8474		0.8740		0.8928
rounded		0.81		0.58		0.74		0.85		0.87		0.89
Rel. Err. Var.		8.5935		25.7806		12.8903		6.4451		5.1561		4.2968
Rel. Std. Err. of M.		2.9315		5.0775		3.5903		2.5387		2.2707		2.0729
Abs. Err. Var.		8.8105		26.4316		13.2158		6.6079		5.2863		4.4053
Abs. Std. Err. of M.		2.9683		5.1412		3.6354		2.5706		2.2992		2.0989

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-03-22 02:25]

INTERNSHIP DATA FROM MENTORS - SOCIAL SCIENCE, 2017/2018 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	1274	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	113497.2781	1273	89.1573	20.0349	20.0349	20.0349	40.7	1.2079
O	548.9393	2	274.4696	0.1926	0.1926	0.1926	0.4	0.1523
PO	73967.7274	2546	29.0525	29.0525	29.0525	29.0525	59.0	0.8140
Total	188013.9448	3821					100%	

G Study Table
(Measurement design P/O)

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% Absolute
P	20.0349		.....		.....	
	.....	O	.....		0.0642	0.7
	.....	PO	9.6842	100.0	9.6842	99.3
Sum of variances	20.0349		9.6842	100%	9.7484	100%
Standard deviation	4.4760		Relative SE: 3.1119		Absolute SE: 3.1222	
Coef_G relative	0.67					
Coef_G absolute	0.67					

Grand mean for levels used: 84.8867
Variance error of the mean for levels used: 0.0875
Standard error of the grand mean: 0.2959

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9922

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	1274	INF	1274	INF	1274	INF	1274	INF	1274	INF	1274	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		3822		1274		2548		5096		6370		7644
Coef_G rel.		0.6741		0.4081		0.5797		0.7339		0.7752		0.8054
rounded		0.67		0.41		0.58		0.73		0.78		0.81
Coef_G abs.		0.6727		0.4066		0.5781		0.7326		0.7740		0.8043
rounded		0.67		0.41		0.58		0.73		0.77		0.80
Rel. Err. Var.		9.6842		29.0525		14.5263		7.2631		5.8105		4.8421
Rel. Std. Err. of M.		3.1119		5.3900		3.8113		2.6950		2.4105		2.2005
Abs. Err. Var.		9.7484		29.2452		14.6226		7.3113		5.8490		4.8742
Abs. Std. Err. of M.		3.1222		5.4079		3.8239		2.7039		2.4185		2.2078

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-04-28 15:31]

INTERNSHIP DATA FROM MENTORS - TECHNICAL EDUCATION, 2017/2018 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	274	INF	
OCCASIONS	O	3	INF	

Analysis of variance

Source	SS	df	MS	Components				
				Random	Mixed	Corrected	%	SE
P	14707.6496	273	53.8742	13.7150	13.7150	13.7150	45.1	1.5528
O	2184.5937	2	1092.2968	3.9400	3.9400	3.9400	13.0	2.8189
PO	6950.0730	546	12.7291	12.7291	12.7291	12.7291	41.9	0.7690
Total	23842.3163	821					100%	

G Study Table (Measurement design P/O)

Source of variance	Differ-entiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% Absolute
P	13.7150		.....		.....	
	.....	O	.....		1.3133	23.6
	.....	PO	4.2430	100.0	4.2430	76.4
Sum of variances	13.7150		4.2430	100%	5.5564	100%
Standard deviation	3.7034		Relative SE: 2.0599		Absolute SE: 2.3572	
Coef_G relative	0.76					
Coef_G absolute	0.71					

Grand mean for levels used: 74.7859
Variance error of the mean for levels used: 1.3789
Standard error of the grand mean: 1.1743

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9912

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	274	INF	274	INF	274	INF	274	INF	274	INF	274	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		822		274		548		1096		1370		1644
Coef_G rel.		0.7637		0.5186		0.6830		0.8117		0.8434		0.8660
rounded		0.76		0.52		0.68		0.81		0.84		0.87
Coef_G abs.		0.7117		0.4514		0.6220		0.7670		0.8045		0.8316
rounded		0.71		0.45		0.62		0.77		0.80		0.83
Rel. Err. Var.		4.2430		12.7291		6.3645		3.1823		2.5458		2.1215
Rel. Std. Err. of M.		2.0599		3.5678		2.5228		1.7839		1.5956		1.4565
Abs. Err. Var.		5.5564		16.6691		8.3345		4.1673		3.3338		2.7782
Abs. Std. Err. of M.		2.3572		4.0828		2.8870		2.0414		1.8259		1.6668

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-04-17 06:15]

INTERNSHIP DATA FROM MENTORS - VOCATIONAL EDUCATION, 2017/2018 ACADEMIC YEAR. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	86	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	5412.6977	85	63.6788	15.5482	15.5482	15.5482	43.8	3.2760
O	541.5426	2	270.7713	2.9504	2.9504	2.9504	8.3	2.2264
PO	2895.7907	170	17.0341	17.0341	17.0341	17.0341	47.9	1.8368
Total	8850.0310	257					100%	

G Study Table (Measurement design P/O)

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% Absolute
P	15.5482		.....		.....	
	.....	O	.....		0.9835	14.8
	.....	PO	5.6780	100.0	5.6780	85.2
Sum of variances	15.5482		5.6780	100%	6.6615	100%
Standard deviation	3.9431		Relative SE: 2.3829		Absolute SE: 2.5810	
Coef_G relative	0.73					
Coef_G absolute	0.70					

Grand mean for levels used: 75.1240
Variance error of the mean for levels used: 1.2303
Standard error of the grand mean: 1.1092

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9898

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	86	INF	86	INF	86	INF	86	INF	86	INF	86	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		258		86		172		344		430		516
Coef_G rel.		0.7325		0.4772		0.6461		0.7850		0.8203		0.8456
rounded		0.73		0.48		0.65		0.78		0.82		0.85
Coef_G abs.		0.7001		0.4376		0.6088		0.7568		0.7955		0.8236
rounded		0.70		0.44		0.61		0.76		0.80		0.82
Rel. Err. Var.		5.6780		17.0341		8.5170		4.2585		3.4068		2.8390
Rel. Std. Err. of M.		2.3829		4.1272		2.9184		2.0636		1.8458		1.6849
Abs. Err. Var.		6.6615		19.9845		9.9922		4.9961		3.9969		3.3307
Abs. Std. Err. of M.		2.5810		4.4704		3.1611		2.2352		1.9992		1.8250

APPENDIX E_4

G and D Study Analyses of Mentors' Results for 2015/2016 to 2017/2018 Academic Years

File C:\Program Files (x86)\EduG - 6.1e\Data\OFF-CAMPUS TEACHING PRACTICE.gen - [2020-07-07 13:59]

INTERNSHIP DATA FROM MENTORS – 2015/2016 TO 2017/2018 ACADEMIC YEARS. DESIGN (P x O)

Observation and Estimation Designs

Facet	Label	Levels	Univ.	Reduction (levels to exclude)
PERSONS	P	9082	INF	
OCCASIONS	O	3	INF	

Analysis of variance

				Components				
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	1064167.3820	9081	117.1861	29.8962	29.8962	29.8962	51.4	0.5876
O	14905.5632	2	7452.7816	0.8176	0.8176	0.8176	1.4	0.5803
PO	499409.1035	18162	27.4975	27.4975	27.4975	27.4975	47.2	0.2885
Total	1578482.0486	27245					100%	

G Study Table (Measurement design P/O)

Source of variance	Differentiation variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	29.8962		.....		.....	
	.....	O	.....		0.2725	2.9
	.....	PO	9.1658	100.0	9.1658	97.1
Sum of variances	29.8962		9.1658	100%	9.4384	100%
Standard deviation	5.4677		Relative SE: 3.0275		Absolute SE: 3.0722	
Coef_G relative	0.77					
Coef_G absolute	0.76					

Grand mean for levels used: 81.0994
Variance error of the mean for levels used: 0.2768
Standard error of the grand mean: 0.5261

Estimate of Phi(lambda)
Cut Score = lambda = 50
Estimate of Phi(lambda) = 0.9906

Optimization

	G-study		Option 1		Option 2		Option 3		Option 4		Option 5	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	9082	INF	9082	INF	9082	INF	9082	INF	9082	INF	9082	INF
O	3	INF	1	INF	2	INF	4	INF	5	INF	6	INF
Observ.		27246		9082		18164		36328		45410		54492
Coef_G rel.		0.7654		0.5209		0.6850		0.8130		0.8446		0.8671
rounded		0.77		0.52		0.68		0.81		0.84		0.87
Coef_G abs.		0.7600		0.5136		0.6786		0.8086		0.8407		0.8637
rounded		0.76		0.51		0.68		0.81		0.84		0.86
Rel. Err. Var.		9.1658		27.4975		13.7487		6.8744		5.4995		4.5829
Rel. Std. Err. of M.		3.0275		5.2438		3.7079		2.6219		2.3451		2.1408
Abs. Err. Var.		9.4384		28.3151		14.1575		7.0788		5.6630		4.7192
Abs. Std. Err. of M.		3.0722		5.3212		3.7626		2.6606		2.3797		2.1724