

Validation of Mentee-Teachers' Assessment Tool within the Framework of Generalisability Theory at the Faculty of Education, University for Development Studies

Simon Alhassan Iddrisu

University for Development Studies,
P. O. Box TL 1350, Tamale, Northern Region, Ghana

ABSTRACT

Practitioners in assessment and other researchers have over the years expressed dissatisfaction with the lack of consistency in scores obtained from use of multiple measurement instruments. Such scores derived from these largely inconsistent and unreliable procedures are relied upon by decision makers in taking very important decisions in education, health and other related fields. The purpose of this study therefore, was to apply the G theory procedures to validate the mentee assessment tool being used at the Faculty of Education, University for Development Studies. The G study involved estimating the generalisability (reliability-like) coefficients of mentees' assessment scores, and determining the level of acceptability (validity) of these coefficients. A nested design was used because different sets of raters assessed different student-mentees on different occasions in the field. The relationship among these variables (facets); students, raters and occasions, appropriately mirrored a nested relationship. Data obtained by raters on 300 students, in the 2018/2019 off-campus teaching practice, were entered into EDUG software for relevant analysis of the results. The study found that both rater and student facets accounted for the largest measurement errors in mentees' observed scores, reporting an estimated G coefficient of 0.62(62%), and representing a positive moderate relationship. Based on these findings, the study concluded that, the quality of mentee observed scores could be improved for either relative or absolute decisions by varying the number of levels of both raters and occasions. To achieve acceptable G coefficient values of 0.83 and above, it is recommended that, decision makers employ a model that uses four raters per occasion for three occasions of assessment.

Keywords: Object of measurement, Relative decision, Absolute decision, universe of generalisation, universe of admissible observations, Composite facet, Generalisability study, Decision(D) Study, Optimisation

INTRODUCTION

Educational assessors and researchers alike increasingly express concern about the reliability and validity of scores produced from multiple measurement procedures such as tests, rating scales, surveys and other forms of observations (1-Alkharusi, 2012). This is because scores generated through the use of any measurement procedure in educational and psychometric assessments, often are basis for making very important decisions (Kolen and Brennan, 2014, 1995; Hughes & Garrett, 2018). Kolen (2014) identified three levels of decision-making based

on assessment scores which are: the individual, institutional, and public policy levels. Individual level decisions based on results may involve a student opting to attend a certain tertiary or non-tertiary institution or even electing to pursue a certain programme of study (Fendler, 2006; Kolen & Brennan, 2014). Institutional level decisions likewise rely on previous assessment records to either certify professionals or to admit applicants into tertiary programmes in relevant institutions. Public policy level decisions address general problems such as improving quality and access to education in the nation for all to benefit from. Shavelson and Webb (2004) submitted that, the usefulness of any assessment score largely is dependent on the extent to which we can generalize with accuracy and precision to wider set of situations.

Allen and Yen (2011) also reckoned that, assessment results generally have multiple purposes and applications, such as found in the selection of new employees, applicants or clients for varied reasons. Yukawa, Gansky, O'Sullivan, Teherani and Fieldman (2020) maintained that in the training of budding professionals in the fields of education, health, law, agriculture, business, etc, assessments remained integral in the process, where relevant rating scales or tools are administered periodically in the conduct of these assessments.

Atilgan (2019), likewise indicated that the choice of a tool for assessment in education depends on what attribute to be measured. Essay-type instruments and tailored rubrics are among several tools reviewed in literature for purposes of assessing the writing skills and other competencies of trainees (Atilgan, 2019; Atilgan, Kan & Aydin, 2017; Turgut & Baykul, 2010). Graham, Harris, and Herbert (2011) for instance, used a writing-based essay type rubric in assessing students' writing skills at the primary. Fleming, House, Hanson, Garbutt, Kroenke Abedin and Rubio (2013) developed the Mentoring Competency Assessment tool (MCA), which they used in assessing the skills of mentors in clinical and translational science. An estimate of the reliability and validity of scores obtained from the MCA tool showed high and positive relationship among the competencies examined.

Many educational settings, schools and other similar institutions are arguably the largest consumers of data emanating from multiple testing and other assessment procedures (Miller et al., 2011). A major challenge associated with measurements in both the social sciences and education is therefore the inconsistencies (unreliability) of its measurements (Sirec, 2017, Revelle, 2016, Brennan, 2005). When the same characteristic is measured on two different occasions, the results obtained often are different (Steyer, 1999; Revelle, 2016). Steyer et al. (1999), also intimated that, irrespective of the measures an institution or body may put in place to assure the sanctity of scores produced from measurements processes, many potential sources of error continue to persist and must be removed.

Many studies cited in literature have used the G theory in investigating the reliability of rating scales and scores obtained from such ratings. Kim, Schatschneider, Wanzek, Gatlin and Otaiba (2017), in a study examined raters and tasks as facets of interest contributing to measurement error in another generalisability study and reported that, the rater facet was a major contributor to measurement error. Sudweeks, Reeve and Bradshaw (2005) similarly, estimated the individual contributions to total variance in a G study with raters and occasions as variables (facets) of interest. However, they reported that, a rater's years of experience in teaching contributed more to measurement error than the rater factor. This report on rater contribution

to measurement errors contradicted that by Kim et al. (2017), who reported a substantial contribution by the rater factor to measurement error.

Researchers like Kan (2007), Graham, Hebert, Sandbank and Harris, (2016), Bouwer, Beguin, Sanders, and van der Ber (2015), and Gebril (2009) variously conducted G studies on a guidance tool rating scale, with the number of essay samples, types of essays, and types of tasks set as factors of interest. Atilgan (2019), examined the reliability of essay rating rubrics using a G theory framework, while Lin and Xiao (2018) investigated the rater reliability using holistic and analytic scoring keys within G theory procedures. With these different studies, the ultimate goal was to quantify the contribution of the individual facets and their composites to total variance.

G theory was chosen over Classical Test Theory (CTT) and Item Response Theory (IRT) due to its superiority in quantifying multiple sources of error in a single study (Brennan, 2005). Whereas the Classical Test Theory focuses on the measurement of reliability in order to differentiate among individuals (Cardinet et al., 2010), G theory in addition, enables the user to evaluate the quality of measurements, not just among individuals but also objects (Cardinet, Johnson, & Pini; 2010). Again, while in CTT, coefficient values determined serve as global indicators of the quality of measurement, G theory does not only calculate the coefficients, but it further provides information on the relative contributions and importance of the different sources of measurement error. Through this unique function, G theory thus permits the user to module the factors into a measurement procedure to improve measurements.

The Generalisability theory practically is applied at two levels, namely; the G study and the decision (D) study levels (Heitman, Kovaleski & Pugh, 2009). Whereas the G study enables the estimation of variance components and reliability coefficients, the D study enables the investigator determine optimal number of levels of facets and possibly, 'positively impact interrater reliability' for making decisions (Moskal & Leydens, 2000, p.28). It allows the user to alternatively employ different levels of the variables involved so as to improve the quality of measurements.

Statement of the Problem

The Faculty of Education, University for Development Studies, trains professional teachers for the various levels of education, in line with the national aims of producing quality teachers. This Faculty of Education introduces new programmes and creates more academic departments. Like many curricula used for training of professionals, that for the training of professional teachers have two main components; namely, the content, and pedagogy (practical) aspects. Pedagogical training, equips students with the relevant professional skills and attitudes they require to enable them teach proficiently in the classrooms, at all levels of education.

The practical components of the training which are implemented as; school observation, On-campus or peer teaching, and Off-campus teaching practice, often are assessed using a tool. This assessment instrument over the years (since 2012/2013 academic year) has been changed for at least two different times on account of the grossly unsatisfactory grades the Professional Education Practice Unit (PEPU) receives on behalf of faculty from raters assigned to evaluate mentees.

Both mentees and faculty have at various times vented their displeasure and suspicions regarding the accuracy, and the quality of grades awarded during professional practice sessions. Similar complaints and expressions of dissatisfaction with the observed scores given to students keep recurring with different cohorts of students at the Faculty of Education. Essentially, the problems over the period have been the glaring inconsistencies associated with scores assigned by different raters to students using this assessment tool. Notwithstanding these dissenting views about the dependency of the observed scores, assessments of mentee work in the field remains an indispensable component of their training. It was therefore incumbent on faculty, lecturers, students and the Professional training unit to devise ways of refining the mentee assessment tool, so as to drastically diminish the deficiencies associated with students' grades, assigned to mentees during field practice. The deficiencies associated with the observed scores essentially were the measurement errors whose sources were unknown and needed to be investigated.

It appears this particular tool has never been scientifically studied either within this university or outside it, to determine the error sources and their magnitudes in the measurements. Possible sources of errors linked with mentee scores may include but not limited to raters, occasions of ratings, mood of raters, items used to examine students among several others (Cook et al., 2018; Atilgan, 2012; Brennan, 2005; Shavelson & Webb, 1991). This study therefore set out with a primary purpose to examine closely the overall reliability of mentee scores assigned by raters on the different occasions using the G theory framework.

Purpose of the Study

The purpose of this study was to validate the mentee assessment tool through estimating the generalisability (reliability) coefficients, and the variance components, and to estimate the optimal levels of the facets for achieving high reliability estimates.

Specifically, this study was guided by the objectives outlined below:

1. to quantify the rater and occasion-effects on the reliability of mentees' observed scores.
2. to explore the effects of composite factors of interest on reliability of the mentees' assessment scores.

Research Questions

This study was guided by the following questions:

1. What amount of variance in mentees' observed scores is attributable to the rater and occasion facets?
2. What effects do the composite facets have on the dependency (reliability) and validity of students' observed scores?

Significance of the Study

The significance of this research finds expression in the meaningfully contribution to improving existing knowledge and practice in the relevant discipline. First the outcomes have the potential to repose confidence and trust in mentees' assessment scores through the study. By estimating the dependency coefficients of the observed scores, the strength of relationship and quality of scores so produced are established enabling decision makers to be well informed on their decisions based on these scores.

Again, decision makers can tell which individual or composite factor/s contribute the least or largest variance to the total variance, so that measures can be taken to reduce the errors and generate high enough coefficients. By subjecting an assessment procedure (tool) through an empirical study, the scores that emanated from the use of that procedure equally were refined through the process and, therefore it is expected that scores would be acceptable to all stakeholders (both students and faculty), as reflecting true attributes of mentees.

Many researchers and other incidental users of the tool and its products, would and should invest greater "faith" in the tool and its outcomes following this validation study. Also, mentees whose performance were assessed using the tool would have known at least the estimated levels of accuracy in their performance leading to some confidence being reposed in the scores.

Delimitation of the Study

This study was restricted in scope naturally by the choice of a named tool and the application of a particular theoretical framework to study it. Participation in this study was restricted to only raters, and mentees who were captured in documents (assessment tool), from which data for the study were extracted. Within the institutions that use this tool, there were many categories of students pursuing other programmes, who equally were excluded from participating in this study by virtue of not being members of the faculty.

In addition, mentee observed scores obtained during the 2018/2019 off-campus internship were also the only admissible data for analysis in this study. Scores of mentees obtained without the use of the assessment tool were ineligible for use in this study.

Limitations of the Study

Limitations of the study entailed the deficiencies associated with the research methodologies applied in this study, which were capable of affecting the generalisability of the findings of this study. Fixing the object of study (assessment tool), nullified all the possible contributions to all composites of the tool, limiting knowledge on much these composites added to total variance. While the nested G study design is less costly, and also usable in multiple situations, it does not enable the maximisation of information about the facets of interest in the study (Cardinet et al., 2010).

G theory is context specific, and so coefficients estimated for a particular population or universe may not be generalised beyond that particular universe. The findings on a given population or conditions for a given study may not be generalised to another population.

LITERATURE REVIEW

Generalisability (G) Theory

Generalizability (G) theory is a statistical body or procedures for evaluating the dependability (or reliability) of behavioural measurements (Cronbach, Glasser, Nanda, & Rajaratnam, cited in Brennan, 2001; Shavelson & Webb, 1992). The concept of dependability, defines the accuracy with which one generalizes from say a student's observed score on an assessment or other measure to an average that student would have received under all possible conditions (Alkharusi, 2012). These possible conditions include the different forms of the instrument, all occasions of rating, or all possible raters. He indicated that the average score is the same as the universe score, and it is synonymous with CTT's concept of true score. G theory thus defines a

score as dependable and valid if it allows accurate inferences to be made about the universe of admissible observations that they are meant to replace (Allal & Cardinet, 1997).

The G theory extends, amplifies and widens the classical reliability theory by identifying, defining and estimating multiple sources of measurement error (Shavelson, Webb, & Rowley; 1992; Alkharusi, 2012). According to Cronbach et al, cited in Orenz (2018), the philosophy upon which G theory is founded maintained that, researchers and assessors are interested in the precision or reliability of a measure because they wish to generalise from the observation in hand to some class of observations to which it belongs (p.144). For instance, teachers would often wish to generalise students' performance on particular occasions over future student performances on multiple occasions.

In G theory, the presumed multitude or infinite number of occasions over which the investigator wishes to generalize over is termed a universe of occasions (Shavelson & Webb, 1991). Likewise, an investigator may want to generalise over a universe of items, raters, situations of observations and students among others. This set of conditions of measurement the researcher might desire to generalise over is the *universe of generalisation* (Alkharusi, 2012; Shavelson, Wiley & Yin, 2015). This universe of generalisation varies from study to study and so a researcher must define it explicitly by identifying the conditions of measurement he/she desires to generalise to in the study (Brennan, 2010).

Again, in G theory language, a set of conditions of measurement of the same type are the facets (Alkharusi, 2012; Brennan, 2010). For instance, the occasions of assessment, raters who conduct assessments, rating scales or items used may constitute facets in a study (Shavelson, Webb & Wiley, 1992; Menendez and Gregori-Giralt, 2018).

Generalisability theory is useful if the desire is to refine the measurement tools/procedures to yield improved and reliable data (Heitman, Kovleski & Pugh, 2009). G theory was chosen over the CTT because of its superior alternative procedures, which yield more useful intra-class correlation coefficients (Denegar & Ball, 1993). G theory further addresses the dependability of measurement issues, and allows the simultaneous estimation of multiple sources of variance, including interactions (Shavelson & Webb; 1991, Naizer, 2007; Morrow, 1989) within a single study. This theory permits a decision maker to investigate the dependability of scores for different kinds of interpretations and uses, making it more preferred choice among other theories for this study.

Generalisability Theory and Validity

Validation is a process of evaluating the claims, assumptions and inferences that link assessment scores with their intended interpretations and uses (Kane, 2006). To validate an assessment tool using G theory procedures requires the merging of the concepts; "reliability" and "validity" (Gugiu, Gugiu, & Baldus, 2012). Orenz (2018), reported another approach to validation which requires the investigator to commence by defining the universe of interests, and then observe two or more independently selected conditions within that universe. A relationship between a generalizability coefficient and validity is thus inferred, where the G coefficient resolves into a validity coefficient. And because the generalisability coefficient indicates how accurately universe scores can be inferred from observed score, it can be interpreted as a validity coefficient (Brennan, 2011).

Contemporary validation frameworks, identified five sources of validity evidence without classifying them as 'types' of validity. For instance, Messick cited in Cook and Beckman (2006), listed five sources of evidence in support of construct validity as; content, response process, internal structure, relations to other variables and consequences. These sources of validity evidence are not types but represent categories of evidence, that can be gathered to support construct validity conclusions derived from assessment scores (Cook & Beckman, 2006). Besides the popular Cronbach's Alpha(α) coefficients which are the generally acceptable standards for interpreting these reliability and validity coefficients, (Naumenko, 2015), George and Mallery (2003) proposed guidelines which served as yardsticks for interpreting validity-like coefficients. These guidelines provide that a coefficient alpha value greater than 0.90 be interpreted as "excellent", around 0.80 as "good", with 0.70 as "acceptable", 0.60 as "questionable", around 0.50 as "poor", and figures below 0.50 as "unacceptable". For purposes of interpreting coefficients in this study, descriptors for professional performance such as *outstanding* shall apply for coefficients 0.90 or higher, *proficient* for coefficients from 0.80 to less than 0.90, *competent* for coefficients from 0.70 to less 0.80, *apprentice* for coefficients from 0.60 to less than 0.70, and *novice* for coefficients from 0.50 and below. These descriptors seem more appropriate for describing performance and categorizing different levels of performance by trainees.

Sources of Measurement Error

Measurement error is an integral part of any assessment exercise, and therefore, it is a truism that every measurement situation or practice is susceptible to some measurement error of some sort (Menendez & Gregori-Giralt; 2018). Measurement errors may manifest as random or systematic types of error (Revelle, 2016; Marcoulides, 1993) as indicated earlier. These measurement errors vary due to differences in raters' inconsistencies in rating, and consequently culminate in a decrease in reliability and validity of the outcomes (Atilgan, 2019).

Tavares, Brydges and Turner (2018) proposed a departure from the use of the traditional classical methods of estimating reliability, to more contemporary procedures like the use of G theory methods which have the advantage of differentiating error sources, and quantifying the contributions of each source to total measurement error. These diverse sources of measurement error may include information introduced through external sources (Sadler, 2009); differences in severity or leniency of scoring by assessors (Knight, 2006), occasions of assessment or differences in the level of training or experience of raters (Howell, 2014). Other error sources may include variations in the depth of meanings or interpretations obtained by raters of the rating scale or tool (Tan & Prosser, 2004).

Generalisability Designs Studies

G theory, according to Brannan (2010), recognizes and distinguishes among designs as either crossed, nested or mixed, and facets as random, fixed, finite or infinite. In a crossed measurement design, all conditions of the facet, for instance, items are observed with all conditions of another source of variation (e.g persons). If raters were crossed with students and occasions, it means all students will be assessed by all raters on all occasions (Shavelson & Webb, 2018;1991). However, in some situations, not all raters observe all students on all occasions. Instead, for reasons of resource constraints, different raters may assess different students on different situations or occasions in a nested design (Cardinet et al., 2010). A notation $i:s$ or $i(s)$ is used to denote facet i nested within facet s . Both conditions stated in the

definition must be upheld for a facet to be nested within another (Cardinet et al., 2010; Shavelson & Webb, 2018).

G theory may also consider a facet(s) as either “random” or fixed in study design. A sample is random when it is drawn randomly or assumed exchangeable with any other sample of the same size obtained from the universe (Gugiu, Baldus & Gugiu, 2012; Cardinet et al., 2010; Shavelson & Webb, 1991). In G theory, the number of variance components produced is dependent on the design adopted for the study. This is because, the variation would be partitioned for the object of measurement, the facets, the interactions, and the residuals.

Advantages of G Theory

The advantages of G theory have been espoused by Shavelson and Webb (1991), Thompson (2003), and Rentz (2018) as follows:

G theory applies less restrictive assumptions, when used in study designs. Particularly, it operates on the assumption that persons and other measurement facets are randomly sampled. Again, unlike other theories, G theory can be used to assess multiple sources of error simultaneously in a single measurement (i.e., items, raters, occasions, etc.). It has the unique property of identifying and estimating separately, each of the possible sources of measurement errors and their combined effects. It is advantageous to an investigator whose measure combines several components in a study of a construct and its dependability. G theory also distinguishes for the researcher, between the magnitude and type of errors, to enable decision makers decide whether the levels of errors are within permissible limits for subsequent applications. G theory also estimates and reports generalisability coefficients explicitly without any ambiguity. It distinguishes between relative and absolute coefficients, which are applied respectively in taking relative and absolute decisions. Other theories like the CTT do not specify the type of coefficient estimated and may not specify the universe to which researchers may generalise to unlike the G theory.

Notwithstanding the obvious advantages G theory has over the other theories, it is not without limitations in its applications (Alkharusi, 2012; Strube, 2002; Shavelson & Webb, 1991; Webb, Rowley, & Shavelson, 1988). First, G theory is context specific, and so coefficients estimated for a particular population or universe cannot be generalised beyond the particular universe. The findings on a given population or conditions for a given study may not be generalised to another population.

Research Design

A partially nested G design was used to estimate and quantify the individual and composite contributions to the total variance in the mentee’s observed score. The G theory procedures were found more appropriate, for they permitted me to isolate, quantify, and to describe the contributions of constituent sources of measurement error. The outputs from G theory analysis relatively were simple and easy to interpret relative to other theories. G theory was thus, justifiably selected for its versatility and for its multiple utility in simultaneously quantifying, and segregating variance components of multiple error sources (Cardinet et al., 2010; Shavelson & Webb, 2005). The design was implemented at three levels; Observation, Estimation and Measurement designs as explained below:

Observation design described the data structure that identified specific facets and their interactions, involving the number of levels of each facet, such as raters, students and occasions (Cardinet et al, 2010). The facets of interest were the assessment tool (Object of measurement), Students (mentees), Occasions and Raters. For the purposes of this study, raters were reorganized into two sets; rater 1s and rater 2s based on which occasion the rater did the assessment. The nested design; $(r:o) \times ps$, represented different sets of raters, who rated different groups of students on different occasions, using the same assessment tool.

The G study was implemented at two levels; involving a generalisability (G) study, and a decision (D) study, (Brennan, 2001; Gugiu, R., Gugiu, C & Baldus, 2012, Marcoulides, 1999). While the G study estimated the G coefficients (G-relative and G-absolute), the D study relied on the information obtained from G study to conduct an optimization study using alternative levels of the factors (facets) of interests specified under the observation design (Cardinet et al., 2010).

The measurement design for this study is denoted by the relationship, SP/RO . A diagrammatic representation of the observation design is shown in Figure 2, originally used by Cronbach et al, (cited in Cardinet et al, 2010). The intersecting circles in Venn diagrams described the regions common to the ellipses (circles) used in variance partition diagram, represented the contributions or "effects" of facets to total score variance.

The nesting of raters, r within occasion, o is represented by the inclusion of one entire circle within another circle.

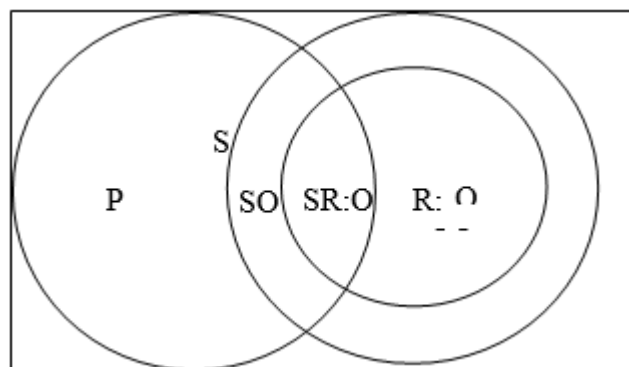


FIGURE 2: Venn Diagram for $ps \times (r:o)$ design

This D study, represented by $ps \times (r:o)$ design, reported five variance components whose estimations were carried out as follows: σ^2_o (occasion variance), was divided by the number of occasions, and the number of raters, $(n_r \ n_o)$ in the D study, by virtue of it being confounded with σ^2_{ro} . In the same way, σ^2_{po} (variance of the interaction between instrument and occasions) was divided by the number of occasions and raters by virtue of its confounding with $\sigma^2_{pro,e}$.

Population

This study targeted all the assessment scores of all raters who assessed mentees at the Faculty of Education. The category of students of interest were those who participated in the 2018/2019 Off-campus teaching practice at both basic and Senior High schools. Out of a total student population of four hundred and thirty (430) who participated in the Off-campus teaching practice, three hundred (300) of them, whose records were available and complete for

the two occasions were used for this study. Table 5 shows the breakdown of students by programme, summing up to three hundred (300).

Table 5: Number of Students by Departments

Serial No.	Programme of Study	Total number of Students
1	Social Science Education	100
2	Business Studies Education	90
3	Mathematics and Science Education	30
4	Agric. and Consumer Sciences Education	20
5	Basic and Early Childhood Care Education	60
	Total	300

Sample and Sampling Procedure

This study used all mentees who had scores for both first and second round of assessments from raters. Criterion sampling procedure (Ary et al., 2014), was employed to sample the 300 mentees who met the criterion for inclusion into the sample. The criterion for selection was that a mentee should have a score from each of the first and second occasion- raters respectively.

The sample size of 300 mentees was determined for this study relying on Krejcie and Morgan (2006) Table of sample size determination which prescribed a sample size of 309, for a given population of 400 at a 3.5% margin of error. Combining the suggested sample size by Morgan's Table and the criterion set for inclusion, and the EDUG requirement for a balanced data for optimum application, I settled on a sample size of 300 mentees who participated in the 2018/2019 off-campus teaching practice programme at the Faculty of Education, University for Development Studies. The adequacy of this sample size is supported by Vispoel, Morris and Kilinc (2018), who used a sample of size of 206 participants, in a G study estimating various reliability coefficients. Atilgan's (2013) proposition that, a sample size of 400 is sufficient for performing an accurate and reliable estimation of the generalisability (G) and Phi coefficients for populations beyond five thousand (5000) also authenticated this sample size for a quantitative G study.

The records of a total of thirty-six (36) raters comprising 15 from Tamale campus, 14 from Wa, and 7 from the Navrongo campus were used for the study. Table 4 contains a summary of raters based on the campus of location.

Table 4: Number of Raters on the Campuses

Serial/Number	Name of Campus	Number of Raters
1	Tamale	15
2	Navrongo	7
3	Wa	14
	Total	36

Data Collection Instruments

The mentee assessment instrument comprised twenty (20) items describing the measures of attributes of teacher-related professional attributes, organized under three sub-dimensions. The items on these sub-dimensions are placed on a five-point Likert-like scale, indicating the

various levels of performance. The points on the scale are 1- 5 with corresponding descriptors as, Novice, Apprentice, Competent, Proficient and Distinguishing/Outstanding. Another column precedes these five columns and it is labelled 'Item and Score'. This column captures the core items the student's performance is being rated against together with a score box. There are three distinct sub-dimensions on the instruments which are; Objectives and Core Points in Lesson Plan, Classroom Organization and Management, and Teaching Methodology and Delivery.

Under the first sub-dimension; Objectives and Core Points in Lesson Plan, respectively are the Objectives, Summaries/Core points, Teaching and Learning activities, Teaching and Learning Materials (TLMs), as well as Subject and Pedagogical Knowledge. These sub-dimensions focus on measuring the extent to which a mentee has mastered these core professional attributes in the lesson plan. The second sub-dimension is Classroom organisation and Management. Three core attributes required by student-teachers include, classroom management and control skills or techniques, and a good professional teacher's attitude in general. Sub-dimension three is "Teaching methodology and delivery" which covers core items 9 to 20. Core items under this sub-dimension include; Introduction of the lesson, presentation of teaching and learning activities, pace of lesson and audibility of voice, questioning and feedback techniques, professional use of chalkboard, use of teaching and learning materials (TLMs), communication skills, Student participation, Assessment for student learning, Mastery of subject matter, Closure of lesson and Professional commitment. These core items in sub-dimension three represent the central activities both learners and teachers, must be engaged in a lesson to achieve learning.

Data Processing and Analysis

Mentee assessment scores which formed the data used for this study were extracted from documents. Access to these records was granted by the Dean of the Faculty of Education, and the Professional Education Practice Unit, who had custody of these records. These records were then sorted, counted and arranged for easy identification and use. To protect the identities and privacies of both student- and rater groups, and also, to avoid the violation of individual rights to secrecy, anonymity and non-intrusion, labels and codes were used in place of their names as identification numbers.

Scores, raters and students' characteristics such as gender, programmes, occasion of assessment among others were extracted from completed assessment tools and the results tabulated under three columns; students, raters and occasions.

Raters were regrouped into two, and designated as rater1s and rater2s depending on whether they assessed students on the first or second occasions. The first column was labeled students with codes from 001 to 300, consisting of the entire group. Student scores accordingly were recorded under rater1 and rater2 respectively in preparation for entry into the EduG software as defined under the observation design.

Data so obtained was then prepared and entered into the EduG software for analysis by research questions posed for the study. Research question one examined major contributors such as raters (r), and occasions, (2) interaction effects; and their residuals to measurement error in mentees' observed score in the G study. Variance components were computed for

individual and composites as: (1) instrument (p), together with other unaccounted for error components. Research question two also examined, estimated and compared the contributions of facet composites to total error variance, with the view to determining the overall effects on the dependency of these mentees' observed scores.

RESULTS

G theory procedures were applied in this study to explore ways of generating dependable scores using the mentee assessment tool, irrespective of the raters or occasions involved when assessments are conducted. This study involved estimating the generalisability (reliability-like) coefficients of mentees' observed scores, computing variance components of facets, and their composites. The information obtained from the G study was then used in subsequent decision studies to explore how G coefficients could be improved to acceptable quality or conventional levels. Following are presentations of varied data groupings relevant to this study:

Background Characteristics of Raters and Students

Raters and students' background characteristics of interest in this study such as sex, campus location and programme specializations or departments were examined, and the statistics captured in Table 9 and 10.

Table 9: Distribution of Raters by Campus and Gender

Campus	Number of Raters by Genders		Total
	Male	Female	
Tamale	12	3	15
Navrongo	7	0	7
Wa	10	4	14
Total	29	07	36

Table 9 displays the distribution of raters by campus and sex as obtained from the records used for this study. The total number of male raters from across the three campuses were twenty-nine (29), while that of female raters were seven. All campuses except Navrongo, had female representation from the list of raters captured in Table 9. Table 10 also displays the distribution of respondents by sex and programme.

Table 10: Distribution of Students by Programme and Sex

Programme/Department	Sex of Student		Total
	Male	Female	
Social Science Education	96	35	131
Business Education	78	13	91
Mathematics and Science Education	36	2	38
Agricultural Science Education	23	1	24
Basic and Early Childhood Education	10	6	16
Total	243	57	300

In Table 10, the students involved in this study were spread across five departments under the Faculty of Education located across the campuses of UDS. Of the total of 300, 131 Social Science students formed the majority, and represented over forty-three (43.6%) percent of the total, ninety-one (91)) students representing just over thirty (30.3%) percent offered Business

Education programme, while sixteen (16) students representing just over five (5.3%) percent of the total pursued Basic and Early Childhood Care Education programme.

Table 11 illustrates the descriptive statistics of raters as presented above.

Table 11: Descriptive Statistics of Rater Scores

	N	Range	Mean	Std. Deviation	Variance
RATER1 SCORES	300	59	70.41	7.975	63.594
RATER2 SCORES	300	53	70.53	8.001	64.009
Valid N (listwise)	300				

In Table 11, the mean scores of raters who assessed students on each of the two occasions are presented. While the mean score reported for rater1s was 70.41, with a standard deviation of 7.98, that of rater2s was 70.53 with corresponding standard deviation value, 8.00 reflecting a very negligible difference between their mean scores. These mean scores obtained for the two pairs of raters were close in size suggesting some similarity in the mode of scoring by these pairs of raters.

Analysis

Analysis of variance (ANOVA) procedure under the G theory was applied to estimate variance components of variables specified in the design as contained in Table 12. The resulting observation and estimation design for the G study are illustrated in Table 12.

Table 12: Observation and Estimation Designs

Facet	Label	Level	Universe
Instrument	P	1	1
Students	S	300	INF
Occasions	O	2	INF
Raters nested within Occasions	R:O	2	INF

In Table 12, G study considered the facets; Students, Occasions and Raters as random facets, because each group of students for instance were drawn from an infinitely large universe of students, and occasions or groups of raters that could equally have participated in this study. Raters and occasions were also denoted random facets, based on similar assumptions. The estimation design in this study thus included three infinite random facets; Students, Occasions and Raters with the assessment tool, designated as a fixed. The measurement design depicting the relationship among the facets in this design was SP/RO.

Research Question 1: What amount of variance is attributable to the rater and occasion facets in the estimated total variance of mentees' observed scores?

Results of the analysis presented the contributions of the multiple sources to the measurement error in the G study as requested by the research question 1, are presented in Table 13. The *ps* (r:o) G design, presented in Table 13, depicted multiple sources of variance components, where the mentee assessment tool (P), and students (S) were the facets of differentiation, and the occasion and rater facets were designated as the instrumentation facets. All estimates of the variance components illustrated in Table 13, were positive except that of the occasion facet

which reported negative variance components (-1.48994), resolved to 0.0 in line with the recommendations of Brennan (cited in Shavelson & Web, 1991).

The contribution of the object of measurement (p) was reported as dots or null values in Table 13, and these represented zero percentage (0.0%) contribution to measurement error. This result aligned well, and in keeping with the G theory principles and procedures which hold that, fixed facets do not contribute to measurement error (Cardinet et al., 2010). The student (S) facet, produced close to 29% ($\sigma^2_s = 41.04$; 28.8%) of the total variance in Table 13, being one single largest contributor to measurement error in this study.

Again, the contributions to total variance by components of the nested relationship, (R:O), was reported as 1.9% ($\sigma^2_{r.ro} = 2.70$). This percentage represented the third largest contribution among the components as reported in Table 13. This result is supported by the findings of Kim et al. (2017), who also found the rater facet as one of the greatest contributors to measurement error. The rater effect (r) was confounded with the rater-by-occasion interaction (ro) and therefore, the variance component for raters (σ^2_r) was hidden in rater-by-occasion interaction (σ^2_{ro}). By convention, the variance component for these confounded effects is denoted as $\sigma^2_{r.ro}$. Compared with the student facet, the raters' contribution to measurement error is considered relatively smaller, and suggests that raters differed minimally in the manner (leniency/harshness) with which they assessed mentees on the different occasions of assessments (Cardinet et al., 2010). It is the desire of every user of an assessment procedure that any error occurring in course of measurement be very small or negligible. This is because the smaller the error, the better the quality, accuracy or precision of the measurement. Decision-makers would be more confident in relying on the scores to make important decisions.

Table 13 : Analysis of variance

Components								
Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P								
S	78996.60	299	264.20	41.044	41.044	41.044	28.8	5.76
O	17.28	1	17.28	-1.49	-1.49	-1.49	0.0	1.07
R:O	1817.27	2	908.63	2.70	2.70	2.70	1.90	2.14
PS								
PO		...						
PR:O		...						
SO	29907.72	299	100.03	1.31	1.31	1.31	0.90	4.9525
SR:O	58250.74	598	97.41	97.41	97.41	97.41	68.40	5.62
PSO								
PSR:O								
Total	168989.60	1199					100%	

Table 13 further illustrated both crossed and nested facet relationships alongside their combined effects, and contributions to total error variance. The paired crossed relationships included; PS, PO, PR:O, SO, and SR:O. In addition, the three-way interaction relationships reported were PSO and PSR:O. The fixed instrument (p) facet had significant effects on other components, like the standard error of measurement (SEM) values (Cardinet et al., 2010). Consequently, the composites of all facets associated with the instrument reported null values

as illustrated in Table 13. These reported effects corroborated the findings of Vispoel et al (2018), and Cardinet et al (2010), who, held the position that fixed facets have nulling effects on all its composites.

The student-occasion interactions (SR:O) nested in occasions reported the highest variance contribution ($\sigma^2_{SR:O}=97.41$, 68.4%) to total variance. This variance also represented the residual component reporting a substantial amount of variation compared with all other composites reported in Table 13. This comparatively larger variance contribution could be explained partly by the confounded facets and the unmeasured sources of variation in this design. Student-Occasion interaction (SO) also reported a nonnegligible variance component of ($\sigma^2_{SO} =1.31$; 0.9% of total variance) implying that, the relative standings of students in respect of the attribute of professional knowledge and skills were unstable on each occasion of assessment of these mentees.

Research Question 2: What effects do the composite facets have on the dependency (reliability) of students' observed scores?

This question was posed purposely to enable me estimate, quantify, and examine the contributions of these facet combinations or interactions to the total variance in the observed scores. In Table 15, all interactions of all other facets with the fixed facet yielded nothing or zero contribution to total variance. Table 14 captured the ANOVA results from which conclusions about the quality and precision of measurements for the selected facets under this measurement design may be drawn.

Table 14 : G Study Table (Measurement design PS/RO)

Source of Variance	Differentiation Variance	Source of Variance	Relative error Variance	% Relative	Absolute error Variance	% Absolute
P	(0.00000)					
S	41.04421					
		O			(0.00000)	0.0
		R:O			0.67602	2.6
PS	(0.00000)					
		PO				
		PR:O				
		SO	0.65414	2.6	0.65414	2.5
		SR:O	24.35232	97.4	24.35232	94.8
		PSO				
		PSR:O				
Sum of variances	41.04421		25.00645	100%		100%
Standard deviation	6.40658		Relative SE: 5.00065		Absolute SE: 5.06779	
Coef-G relative	0.62					
Coef-G absolute	0.62					

In both Tables 14 and 15, variance components for each of the facets of generalisation are different. The tool was fixed to reduce random fluctuations and as such, it made no

contributions to measurement error (Cardinet et al., 2010 & Shavelson and Web; 1991). The student facet was not fixed, and so contributed to true variance ($\sigma^2_s = 41.04$: 28.8%) accounting for as much as 28.8% of total variance as shown in Table 14.

The composite facets, whose effects were being examined included, PS, PO, PR:O, SO, SR:O, PSO and PSR:O (Table 15). Among these composites, only two made significant contributions to total variance, namely; SO and SR:O. The zero effect of the fixed facet (p), was amply manifested in the interactions that follow; PS, PO, PR:O, PSO and PSR:O, registering null values in each of these cases (Cardinet et al., 2010; Brennan, 2005). The interaction between students and occasions (SO) was thus, the only pair that made a nonnegligible contribution to absolute error variance in the measurement design, accounting for ($\sigma^2_{so} = 0.65$, 2.5% of total variance). In comparison with other composite facets, this was a substantial and no mean contribution to variation of student behaviour or performance on the different occasions of assessment, consistent with findings by Sudweeks, Reeve and Bradshaw (2005).

Again, in Table 15, the largest contributor to total variance was the composite between student and rater facets nested in the occasion facet, which reported over 94% ($\sigma^2_{SR:O} = 24.35$, 94.8% of total variance) contribution to variability. The effects of multiple facets such as SR, RO and other unmeasured sources were confounded in this composite variance reported here.

DISCUSSION OF RESULTS

This study set out to examine the quality of mentee scores obtained using an assessment tool through G theory procedures. One key objective of the study was to estimate, and isolate the contributions of facets of interest to total variance in the mentee's observed scores. In this generalisability analysis, variance components, and the G coefficients, involving the relative and absolute coefficients, were used as indicators of the quality of performance or scores. While the G coefficients provided global measure of the reliability of student scores, the variance components mirrored the individual contributions of facets to total measurement error (Cardinet et al., 2010). Information generated through the G analysis served as evidence to support assumptions and inferences derived from these scores. The study used a nested design where the object of measurement, the assessment tool was fixed, and then students, raters and occasions designated as random facets. This result appears consistent with findings reported in many validation studies conducted using G theory procedures in literature (Cardinet et al., 2010; Atilgan, 2019; Naumenko, 2015).

Regarding the individual contributions of rater and occasion facets in response to the first research objective, the study reported a zero contribution by the occasion facet to absolute error variance in line with the findings of Medvedev et al. (2018). The zero contribution by the occasion facet is further corroborated by findings of Sudweeks, Reeve and Bradshaw (2005), which maintained that alternative variables other than just the occasion would add to measurement error. They mentioned variables such as the environment, experience of the rater, quality of the measure among possible sources of error. Again, since the rater facet was nested in occasions, no lone contribution by the rater facet was possible from the nested relationship because of confounding. However, the contribution of the nested relationship of raters and occasions (R:O) to total variance was reported as 2.6% of absolute error variance, representing a relatively small part of the total variance. Since the rater effect was confounded with the rater-by-occasion interaction (ro), the lone contribution to measurement error by

rater variable was not observed in the rater-by-occasion interaction. The percentage thus reported signified a minimal difference in rating performance by raters on each occasion of assessment. The contribution means that, raters either differed minimally in their leniency or severity with which they scored the mentees in the classrooms. Since no main rater effect was reported in this result due to confounding, with a null or zero contribution by the occasion facet, it is only reasonable and logical to infer that the 2.6% contribution to total variance was accounted for by raters in the nested relation. This contribution to total measurement error in the mentee's observed score is supported by the findings by Kim et al. (2017) and Naumenko (2015) as reported in literature.

Objective two of this study also examined the contributions of composite facets to measurement error in mentee observed scores. In Table 14, only two composites made significant contributions to the measurement error besides that of the nested relationship. It was revealed that, the composite facet; Students and Occasions (SO) contributed 2.5% (Table 15) to absolute error variance. This percentage contribution ranks the third highest after the rater nested in occasion(R:O) composite which recorded 2.6% addition to the absolute error variance. The absolute variance components were reported here because the focus was to estimate the precisions of observed scores independently, and to determine the exact position of each individual on the assessment scale.

Again, the student-by-Occasion contribution of 2.5% to absolute variance meant that, on each occasion of assessment, students' 'real' standing on the construct of interest varied by about 2.5%. Ordinarily, an individual's true standing or performance should be repeatable to achieve stability on a construct with time, but a situation where the absolute position kept changing on different occasions provided fertile grounds for suspicious reliability and validity conclusions on such results. This position was further aggravated by the comparatively larger percentage contribution to absolute error variance reported for the student-by-Rater interaction nested in Occasions (SR:O). This composite facet registered the highest effect on absolute error variance accounting for over 94% of total variance in the observed score. This high percentage of absolute variance generated in respect of student-rater interaction signified that raters' ratings of students' performance on the scale remained relatively high on different occasions. If raters scored students high or low scores on the construct of interest which was the student's professional knowledge and skills on one occasion, a similar observation would have been repeated on another occasion of assessment. Alternatively, if raters were severe in rating students on one occasion, they likely might behave same on a different occasion. This inference from the foregoing aligns well with Messick's position cited in Brennan (2005), which identified two categories of raters, namely – a lenient and a strict scorer. By design some scorers award high scores while others also give low scores irrespective of the occasion. Such groups of scorers whether lenient or strict still exhibit subjectivity in their ratings, thereby rendering their observed scores error infested.

SUMMARY OF KEY FINDINGS

Findings on Research Question 1

Based on the results and analysis of data in this study, it was revealed that while the occasion facet made zero contribution to measurement error, the rater facet (r), though confounded (R:O), contributed 1.9%, to total variance in the observed mentee scores. The study also found that, the overall dependency (reliability) and precision (quality) with which the mentee tool

measured the construct of interest in mentees was 62% of the universe (Coef G = 0.62) or the mentees' true score. This measure represented a positive coefficient of moderate strength. In addition, the nested relationship, (R:O) accounted for 1.9% of total variance, which also represented the third largest contributor to total variance among the variance components. Rater facet was found a higher contributor to total variance, than the occasion facet.

Findings on Research Question 2

2). The study also found that, the student and occasion composite (SO), and the student-and-Rater interaction nested in the occasion facet (SR:O), were the only two that made substantial contributions to total variance relative to other composites in the design. The interaction between student and occasion facet (SO) made significant and a nonnegligible contribution to absolute error variance, accounting for 2.5% ($\sigma^2_{so} = 0.65$, 2.5%) of total variance.

The Student and Rater composite, nested in the occasion facet (SR:O), were the largest contributor to absolute variance reporting over 94% ($\sigma^2_{SR:O} = 24.35$, 94.8% of total variance) to variability. The interaction between students and raters on different occasions of assessment made the most substantial contribution to total variance compared with other composites.

CONCLUSIONS

Overall, the study ascertained that, the dependency (reliability) of mentee observed scores as estimated was 0.62 (i.e., 62%), and represents a positive moderate quality level when placed on an absolute measurement scale. These G coefficients represented global measures of the reliability or dependency and precision with which mentees' attributes were estimated. Conventionally, dependency coefficients above 0.80 are more desirable, and preferable for absolute type of decisions. The rater nested in occasion (R:O) composite, the crossed relationship between students and Occasions, as well as the composite between student and rater facets nested in the occasion facet (SR:O) respectively made contributions to total variance in the design. Based on the nonnegligible contribution to total variance by rater-occasion composite, it is instructive that the rater main effect be investigated further.

The students and raters composite comparatively made larger variance contribution than the other composites which meant that, different raters while using the same assessment tool on students, differed substantially in what they observed about student performance on the different occasions. These different raters were either lenient in scoring students, and thus saw more of the construct of interest or were probably harsh in scoring students on different occasions, and therefore observed less of the construct in students. The suggested variations in rater interpretations of items on the measurement tool has implications for both the reliability and validity of scores obtained from these raters. It is therefore imperative that, relevant authority take appropriate steps targeted at ensuring that raters are given adequate exposure and training to enable them develop a common understanding of the tool.

RECOMMENDATIONS

Recommendations made based on the findings and conclusions arrived at follow:

1. Based on finding that, raters and students were the largest contributors to total measurement error in mentees' observed scores, it is recommended that decision makers such as Teaching Practice officials across universities and colleges of education, regularly build the capacities of both raters and students, through periodic tailored

trainings on the contents of the assessment tool so as to reduce errors in rating, and also to clarify student understandings of tool. Periodic intensive trainings are necessary particularly for newly recruited teaching staff who might be using the tool for the first time for mentee assessments.

2. This study used a nested design with the object of measurement fixed, resulting in an extensive nulling effect on all composites of the object of measurement (mentee tool). It is recommended that a crossed design be used in future without fixing the object of measurement to maximize information regarding all facets' contributions to total variance. Alternatively, a trial of the nested design could be used without fixing the mentee tool to explore the individual and composite facet contributions to total variance of the mentee observed scores. A decision maker would be better informed on which design to employ considering which of them is cost-effective, and at the same producing high-quality outcomes.
3. This study solely used G theory in investigating this mentee assessment tool leading to the results and conclusions arrived at. It is possible the use of an alternative theory or procedure could assist in ascertaining and confirming the findings of this study. It is therefore recommended that; a future validation study adopt either factor analysis (FA) or Item Response Theory procedures as a form of confirmatory study of findings of this study.

SUGGESTIONS FOR FURTHER RESEARCH

The following suggestions were made for future researchers who have interest in conducting similar studies:

1. Future researchers who wish to use the G theory for a similar study may still explore the nested design without fixing the object of measurement. Alternatively, such researchers should apply factor analysis or IRT procedures which have the advantage of computing item characteristics besides segregating the measurement errors associated with individual factors of interest.
2. It is recommended that future researchers define the sub-dimensions of the tool as part of the facets, in addition to raters, students and occasions, to examine their separate contributions to total variance in the observed score, under either a crossed or nested design. A revised model which integrates the additional facets could lead to an improved quality of reliability coefficients of mentee observed scores.
3. Furthermore, researchers in future may also consider setting up both crossed and nested designs in one study using separate student groups to compare the G coefficients and error variances in either designs to inform decision making regarding which design is more economical to use in future assignment of raters for assessment exercises.

References

- Adom, D., Husein, E. K., & Agyem, J. A. (2018). Theoretical and conceptual framework: Mandatory ingredients of a quality research. *International journal of scientific research*. <https://www.researchgate.net/publication/322204153>
- Alexopoulos, D.S. (2007). Classical test theory. In N.J. Salkind (Ed.), *Encyclopedia of measurement and statistics*. Thousand oaks: CA: Sage publications, 140-143.
- Allal, L., & Cardinet, J. (1997). Generalisability theory. In J.P. Keres (Ed.), *Educational research, methodology, and measurement: An international handbook* (2nd, p. 737 -741). Cambridge, United Kingdom: Cambridge University.

- Allen, M. J. & Yen, W. M. (2002). *Introduction to measurement theory*. Illinois: Waveland press, Inc.
- Ali, A. M., & Yusof, H. (2011). Quality in qualitative studies: the case of validity, reliability and generalisability. *Issues in social and environmental accounting*, 5(1), p.25 -64.
- Alkharusi, H. (2012). Generalisability theory: An analysis of variance approach to measurement problems in educational assessment. *Journal of studies in education*. <https://www.researchgate.net/publication/256349585>
- Ary, D., Jacobs, L. C., Sorensen, C. K., & Walker, D. A. (2014). *Introduction to research in education*. Wadsworth, Cengage Learning Inc.
- Atilgan, H. (2019). Reliability of essay ratings: A study on generalisability theory. *Eurasian journal of educational research*, 1, 133 – 150.
- Atilgan, H. (2008). Using generalisability theory to assess the score reliability of the special ability selection examinations for music education programmes in higher education. *International Journal of research & methods in education*, 31 (1), 63 –76 <https://doi.org/10.1080/17437270801919925>.
- Baartman, L. K.J., Bastiaenns, T. J., Kirschner, P. A., van der Vleuten, C. P. M. (2007). Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. *Educational research review*; www.elsevier.com/locate/EDUREV
- Babie, E. (2013). *The practice of social research* (13th ed.). Canada: Wadsworth, Cengage Learning.
- Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). Washington, DC. Eric publications.
- Bluman, A. G. (2012). *Elementary statistics: A step-by-step approach* (8th ed.). USA, McGraw-Hill.
- Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model; Fundamental measurement in the human sciences* (3rd ed.) New York, Routedledge group.
- Bouwer, R., Beguin, A., Sanders, T., & van den Berg, H. (2015). Effect of genre on the generalisability of writing scores. *Language testing*, 32 (1), 83 – 100.
- Brannan, R. L. (2013). Commentary on ‘validating the interpretations and uses of test scores. *Journal of educational measurement*, 50: 74-83.
- Brennan, R. L. (2011). Generalisability theory and classical test theory. Centre for advanced studies in measurement and assessment, University of Iowa. *Applied measurement in education*, 24: 1-21.
- Brennan, R.L. (1996). Generalisability of performance assessments. In *technical issues in large-scale performance assessment*, edited by Philips, G., 19 -58. Washington, DC: National center for educational statistics.
- Brennan, R. L. (2003). Coefficients and indices in Generalisability theory. Center for advanced studies in measurement and assessment (CASMA) research report (1).
- Brennan, R.L. (2001). *Generalisability theory*. New York: Springer-Verlag New York, Inc.
- Brennan, R. L. (1998). A perspective on the history of generalizability theory. *Educational measurement: issues and practice*, 16(4), 14-20
- Brennan, R.L. (1996). Generalisability of performance assessments. *Technical issues in large-scale performance assessment*. Washington, DC: National center for education and statistics.
- Brennan, R.L. (1992). *Elements of generalisability* (2nd ed.) Iowa city. ACT publications.

Cardinet, J., Johnson, S., & Pini, G. (2010). Applying generalisability theory using Edu G. Quantitative methodology series. New York, Routledge Group.

Cook, D.A., Thompson, W.G., & Thomas, K.G. (2014). Test-enhanced web-based learning: optimizing the number of questions (a randomized crossover trial). *Academic med*: 169 -175

Cook, D.A., Beckman, T. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *Journal of education*, 19(7), 166.

Crocker, L., & Algina, J. (2008). Introduction to classical and modern test theory. OH, Cengage learning.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Belmont, CA: Wadsworth group/Thomson learning.

Cronbach, J.L. (1984). Essentials of psychological testing. New York: Happers & Row publishers.

Cronbach, J. L., Lin, R. L., Brennan, R.L. & Haertel, E. (1997). Generalisability analysis for performance assessment of student achievement or school effectiveness. *Educational and psychological measurement*, 57, 373 -399.

Denegar, C. R., & Ball, D. W. (1993). Assessing reliability and precision of measurement: an introduction to intra class correlation and standard error of measurement. *Journal sport rehabil*, 2 (1),35- 42.

DeVillis, R.F. (2017). Classical test theory. *Med care*, 44(11), 50 – 59.

Dogan, C.D., & Uluman, M. (2017). A comparison of rubrics and graded category rating scales with various methods regarding raters' reliability. *Educational sciences: Theory & practice*,7, 631 -651. <https://dx.doi.org/10.12738/estp.2017.2.0321>.

Doowning, S.M. (2003). Validity on the meaningful interpretation of assessment data. *Med education*, 37: 830-837.

Doowning, S.M. & Haladya, T.M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ*, 38: 327 -333.

DeShon, R. P. (2002). Generalisability theory. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analysing behaviour in organisations*, 189 -220.

Edenborough, R. (1999). Using psychometrics, a practical guide to testing and assessment (2nd ed.). London, Kogan page Ltd.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R.L. Linn (Ed.), *Educational measurement (3rd ed., pp. 127 -144)*. New York: Macmillan.

Fleming, M., House, S., Hanson, V. S., Yu, L., Garbutt, J., McGee, R., Kroenke, K., Abedin, Z., & Rubio, D. M (2013). The mentoring competency assessment: Validation of a new instrument to evaluate skills of research mentors. *Academic medicine*, 88(7).

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2015). How to design and evaluate research in education (9th ed.). New York, McGraw-Hill education.

Fulton, S. & Krainovich-Miller, B. (2010). Gathering and appraising the literature in LoBiondo-Wood, G. & Haber, J. (Eds). *Nursing research: methods and critical appraisal for evidence-based practice (7th ed.)*. St, Luis MO: Mosby Elsevier.

Gaur, A. & Gaur, S.S. (2009). Statistical methods for practice in research: A guide to data analysis using SPSS. New Delhi, Sage publications Inc.

Gebriel, A. (2009). Score generalisability of academic writing tasks: does one test method fit it all? *Language testing*, 26(4):507 – 531.

Ghana Education Service Headteacher's handbook.

Ghazali, N. H. (2016). A reliability and validity of an instrument to evaluate the school-based assessment system: A pilot study. *International journal of evaluation and research in education (IJERE)*, 5 (2), 148 – 157.

Gierl, M. J., & Bisanz, J. (2001). Item response theory for psychologists -Book review. *Applied psychological measurement*, 25 (4), 405 -408.

Graham, S., Harris, K., & Herbert, M. (2011). *Informing writing: The benefits of formative assessment. A carnegie corporation time to act report*. Washington, DC: Alliance for excellent education.

Grant, C. & Osandoo, A. (2014). Understanding, selecting and integrating a theoretical framework in dissertation research. Creating the blueprint for 'House'. *Administrative issues journal: connecting education, practice and research*. pp.12 -22.

Gullo, D. F. (2005). *Understanding assessment and evaluation in early childhood education* (2nd ed.). New York, Teachers college press.

Gugiu, P.C. (2011). Summative confidence. Unpublished doctoral dissertation, Western Michigan university, Kalamazoo.

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

Hafner, J. C., & Hafner, P. M. (2003). Quantitative analysis of the rubric as an assessment tool: An empirical study of student peer-group rating. *International journal of science education*, 25 (12), 1509 -1528.

Hambleton, R.K. & Jones, R.W. (1997). Comparison of classical test theory and item response theory and their applications to test development.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury; Sage publications.

Hambleton, R.K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development.

Hamrick, L.R., Haney, A. B., Kelleher, B. L., & Lane, S. P. (2020). Using generalisability theory to evaluate the comparative reliability of developmental measures in neurogenic syndrome and low-risk populations. *Journal of neurodevelopmental disorders*, 12 (16).

Heitman, R. J., Kovaleski, J.E., & Pugh, S. F. (2009). Application of generalizability theory in estimating the reliability of ankle-complex laxity measurement. *National Athletic trainers' Association, Inc. Journal of Athletic training*, 44(1), 48 -52.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Boston, MA: Allyn and Bacon.

Howell, R. (2014). Grading rubrics: Hoopla or help? *Innovations in education and teaching international*. 51(4); 400 -410.

Ilgen, J.S., Hatala, R., & Cook, D.A. (2015). A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med education*, 161 -173.

Institute for Continuing Education and Interdisciplinary Research (ICEIR) (2012). University for Development Studies mentoring policy. Tamale, M-bukConcepts publication.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics; reliability, validity and educational consequences. *Educational research review* 2. 130 – 144.

Kan, A. (2007). Effects of using a scoring guide on essay scores: Generalisability theory. *Perceptual and motor skills*, 105, 891 -905.<https://doi.org/10.2466/pms.105.3.891-905>.

Kane, M.T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17 -64). Westport, CT: American Council on Education /Praeger.

Kane, M.T. (2001). Current concerns in validity theory. *Journal of education measurement*, 38: 319 -342.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological bulletin*, 112: 527 – 535.

Kaplan, R. M. & Saccuzzo, D.P. (1989). *Psychological testing: Principles, applications and issues* (2nd ed.). California, Brooks/Cole publishing Co.

Kovaleski, J.E, Hollis, J. M., Heitman, R. J., Gurchiek, L. R., Pearsall, A. W, (IV). Assessment of ankle-subtalar-joint complex laxity using an instrumented ankle arthrometer: an experimental cadaveric investigation. *Journal of athl train*, 37(4), 467 -474.

Kean, J. & Reilly, J. (2014). Classical test theory.192-194

Kim, Y. S. G., Schatschneider, C., Wanzek, J., Gatlin, B., Otaiba, S. (2017). Writing evaluation: rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Read writ*, 30, 1287- 1310.

Kline, J. B. (2014). *Classical test theory: Assumptions, equations, limitations, and item analyses*. Thousand Oaks, Sage publications.

Knight, P. (2006). The local practices of assessment; *Assessment and evaluation in higher education*.31(4);435 - 452.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating methods and practices*. New York, Springer – Verlag Inc.
Lawson, J.E. & Cruz, R. A (2017). "Evaluating special educators' classroom performance: Does rater "type" matter?", assessment for effective intervention.

Leedy, P. O., & Ormrod, J. E. (2010). *Practical research, planning and design* (9th ed.). New Jersey, Pearson education Inc.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational researcher*, 36, 437 - 448.

Mafla, A. C., Herrera-Lopez, H.M., & Villalobos-Galvis, F. H. (2018). Psychometric approach of the revised illness perception questionnaire for oral health (IPQ-R-OH) in patients with periodontal disease. *Journal of periodontology*. DOI: 10.1002/JPER.18-0136.

Magis, D. (2014). Accuracy of asymptotic standard errors of the maximum and weighted likelihood estimators of proficiency levels with short tests. *Applied psychological measurement*, 38, 105 -121.

Marcoulides, G., & Ing, M. (2013). The use of generalisability theory in language assessment. In A. Kunnan, (Ed.), *The companion language assessment*, 3, 1207 -1223. New York, NY: John Wiley & Sons, Inc. DOI: 10.1002/9781118411360.wbcla014

Marcoulides, G. A (1999). Generalisability theory: Picking up where the Rasch IRT models leaves off? In. Embretson, S & Hershberger, S. The new rules of measurement: What every psychologist and educator should know. Lawrence Erlbaum associates Inc., 129 -130.

Marcoulides, G. A. (1996) "Estimating variance components in generalizability theory: The covariance structure analysis approach", structural equation modeling: *A multidisciplinary journal*.

Marcoulides, G. A. (1993). Maximizing power in generalizability studies under budget constraints, *Journal of educational statistics*, 18(2), 197-206.

McDonald, R. P. (1999). Test theory. A unified treatment Mahwah, NJ: Lawrence Erlbaum Associates.

Medvedev, O.N., Merry, A.F., Skilton, C., Gargiulo, D. A., Mitchell, S. J., & Weller, J. M. (2019). Examining reliability of WHOBARS: a tool to measure the quality of administration of WHO surgical safety checklist using generalisability theory with surgical teams from three New Zealand hospitals. *BMJ Open* 2019;9:e022625:doi:10.1136/bmjopen - 2018-022625.

Messick, S. (1995). Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychological Association; 50:741 – 749.

Messick, S. (1989). Validity in educational measurement (3rd ed). New York; American Council on education and Macmillan.

Moskal, B., & Leydens, J. (2000). Scoring rubric development: Validity and reliability. *Practical assessment, research and evaluation*, 7(10), 1-11.

Muijs, D. (2011). Doing quantitative research in education with SPSS. London: Sage Publications limited.

Naizer, G.(2007). Basic concepts in generalisability theory: a more powerful approach to evaluating reliability. <http://www.eric.ed.gov>. Eric document reproduction service no. ED341729. Accessed

Nitko, A. J., & Brookert, S. M. (2011). Educational assessment of student. Boston, MA: Pearson Education.

Nitko, A. J. (2004). Educational assessment of students (4th ed.). New Jersey, Pearson Education, Inc.

Nunnally, J., & Bernstein, I. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill, Inc.

Ojerinde, D., Popoola, K., Oje, F., & Onyeneho, P. (2012). Introduction to item response theory (2nd ed.). Abuja. Marvelouse mike press Ltd.

Orentz, J.O. (2018). Generalisability theory: A comprehensive method for assessing and improving the dependability of marketing measures. *Journal of marketing research*, 24, 19 -28.

Pallant, J. (2007). Step-by-step guide to data analysis using version 15: SPSS survival manual (3rd). McGraw-Hill, open university press.

Regan, B.G., & Kang, M. (2005). Reliability: issues and concerns. *AthlTher today*, 10(6), 30 -33. individual differences, 103, 32-39. Doi: 10.1016/j.paid.2016.04.007.

Revelle, W. (2013). Personality project. Retrieved October, 2013, from <http://personality-project.org/revelle/syllabi/405.old.syllabus.html>

Sadler, D. (2009). "Indeterminacy in the use of preset criteria for assessment and grading". Assessment and evaluation in higher education, 34 (2): 159 -165.

Sadler, P. M. & Good, E. (2006). The impact of self –and peer-grading on student learning. *Educational assessment*, 11(1), 1 -31.

Shavelson, R. J. & Webb, N.M. (2005). Generalisability theory. *British journal of mathematical and statistical psychology*, 59- 612.

Shavelson, R.J, Webb, N., & Rowley, G. (1992). Generalisability theory. <https://www.researchgate.net/publication/344083148>

Shavelson, R.J., & Webb, N.M. (1991). Generalizability theory: A primer. Newbury Park, Sage publications.

Sinclair, M. (2007). "A guide to understanding theoretical and conceptual frameworks." *Evidence-Based midwifery*, 5 (2), 39. *Gale onefile: Health and medicine*. link.gale.com/apps/doc/A167108906/HRCA?u=anon~a99d7734&sid=googleScholar&xid=20965ff (Accessed 5 May 2022).

Sireci, S. (2007). On validity theory and test validation. *Educational researcher*, 36(8), 477 -481.

Sireci, S. G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualization of validity. *Educational measurement: Issues and practice*, 25 (3), 27 -34.

Stangor, C. (2006). *Research methods for the behavioural sciences* (3rd ed). Boston: Houghton Mifflin.

Stegner, A.J., Tobar, D. A., & Kane, M. T. (1999). Generalizability of change scores on the body awareness scale. *Measurement in physical education and exercise science*.

Stemler, S. E. (2007). Cohen's kappa. In N.J. Salkind (Ed.), *Encyclopedia of measurement and statistics*. Thousand Oaks, CA: Sage publications.

Steyer, R. (1999). *Classical (Psychometric) test theory*. Jena, Germany.

Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state–trait theory and research in personality and individual differences. *European journal of Personality*, 13(5), 389 –408.

Sudweeks, R., Reeve, S. & Bradshaw, W. (2005). A comparison of G theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing writing Vol. 9*, 239-261.

Suen, H.K. & Lei, P. W. (2007). Classical versus generalizability theory of measurement. *Journal of Educational measurement*, 4, 1-13.

Schoonen, R. (2012). The validity and generalisability of writing scores. The effect of rater, task and language. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. van den Berg (Ed.), *Measuring writing: Recent insights into theory, methodology and practice (pp. 1- 22)*. Leiden, The Netherlands: Brill.

Stevens, D. D., & Levi, A. J. (2005). *Introduction to rubrics*. Sterling, VA: Stylus publishing.

Strube, M.J. (2002). Reliability and generalisability theory. In L.G grimme & P.R. Yarnold (Eds.) *Reading and understanding more multivariate statistics (pp. 23 -66)*. Washington, DC: American Psychological Association.

Swiss Society for Research in Education Working Group (2010). *Quality of measurement in education- EDUG user guide*. Switzerland.

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics* (5th ed.). Allyn and Bacon/Pearson education.

Tan, K. & Prosser, M. (2004). Qualitatively different ways of differentiating student achievement: A phenomenographic study of academic's conceptions of grade descriptors. *Assessment and evaluation in higher education*, 29(3), 267 -282.

- Tavares, W., Bridges, R. & Tuner, L. (2018). Applying Kane's validity framework to a simulation- based assessment of clinical competence. *Advances in health science education*. <https://www.researchgate.net/publication/320679278>
- Templin, J. (2012). Item response theory. Colorado, Georgia, USA.
- Thompson, B. (2003). A brief introduction to generalisability theory. In B. Thompson (Eds.) *Score reliability: Contemporary thinking on reliability issues*, 43 -58. Thousand Oaks, C.A: Sage.
- Turgut, M. & Baykul, Y. (2010). *Measurement and evaluation in education (Egitimde olome ve degerlendirme)*. Ankara: Pegem Akademi. University for Development Studies Administrative Manual, (2018). University for Development Studies Statutes, (2017).
- Vispoel, W. P., Moris, C. A., & Kilinc, M. (2018). Applications of generalisability theory and their relations to classical test theory and structural equation modeling. *American psychological association; Psychological methods*, 23(1),1-26.
- Webb, N.& Shavelson, R.J. (2005). Generalisability theory: Overview. <https://www.researchgate.net/publication/22758011>
- Webb, N.M., Rowley, G.L., & Shavelson, R.J. (1988). Sing generalisability theory in counselling and development. *Measurement and evaluation in counselling and development*, 21, 81-90.
- Yukawa, M., Gansky, S.A., O'Sullivan, P., Teherani, A., & Fieldman, M.D. (2020). A new mentor evaluation tool: Evidence of validity. *Plos one* 15 (6): e0234345. <https://doi.org/10.1371/journal.pone.0234345>. Accessed on June 16,2020.
- Zanon, C., Claudio, S. H., Hanwook, Y., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexaoe critica*.29:18. DOI 10,186/54:1155 – 016 -0040 -x.
- Zhang, B., Johnston, L., & Kilinc, G. (2008). Assessing the reliability of self-and peer rating in student group work. *Assessing and evaluation in higher education*,33(3), 329– 340.
- Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of modern applied statistical methods*, 6(1), 21 -29.