African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

RELIABILITY OF STUDENTS' TEACHING PRACTICE SCORES USING GENERALIZABILITY THEORY.

Oduro-Okyireh, G.[1], Asamoah-Gyimah, K.[2], Oduro-Okyireh, T.[3], and Nugba, R. M.[4]
[1]Department of Interdisciplinary Studies, Akenten Appiah-Menka University of Skills Training and Entrepreneurial Development (AAMUSTED), Ghana.
[2,4]Department of Education and Psychology, University of Cape Coast (UCC), Ghana.
[3]Department of Mathematics and Statistics, Cape Coast Technical University, Ghana.
[1]*gookyireh@aamusted.edu.gh*
[2]*kasamoah-gyimah@ucc.edu.gh*
[3]*theodoke@yahoo.com*
[4]*regina.nugba@ucc.edu.gh*

ABSTRACT
Purpose: The aim was to demonstrate the potency of using Generalisability Theory (GT) over Classical Test Theory (CTT) in reliability estimation. The objectives are to find the major source of error in the Intern Teaching Evaluation Form (ITEF) scores and to identify the optimum number of occasions of rating of teaching practice that would give the most reliable scores.
Design/ Methodology/ Approach: A random effects one-facet fully crossed design in which intern (p) was fully crossed with occasion (o) was adopted for the study. In all, 9,082 bachelor's degree teaching practice triplicate scores for three academic years from 2015/2016 to 2017/2018 were analysed in this study. A univariate generalisability analysis with EduG was performed to analyse data.
Findings*:* The finding for relative interpretation, the scores were strongly reliable. For absolute interpretation, the scores were moderate to strongly dependable. The major source of error in the ITEF scores was the p x o interaction combined with unidentified sources.
Research Limitation/Implications: A firm conclusion on one major source of error being a single facet in the ITEF scores could not be made.
Practical Implication: The level of reliability and the optimum scoring design of the ITEF established by this study will aid in selecting the ITEF by teacher training institutions in Ghana for evaluation of teaching practice to train pre-service teachers for effective curriculum implementation to achieve national educational aims.
Social Implication: Reliable evaluation data can inform policy decisions related to teacher certification, professional development, and educational standards. Policymakers can use this data to develop evidence-based policies that support teacher quality and student achievement.
Originality/ Value: It has unearthed the psychometric properties and the ideal scoring design of the ITEF which were hitherto unknown.
Keywords*: Evaluation. generalisability theory. students. teaching practice. training institutions.*

INTRODUCTION

ISSN: 2408-7920

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

Observational techniques in assessment are widely used in the Ghanaian education system for the evaluation of teaching skills by the 46 Colleges of Education and other accredited teacher training universities in Ghana (Quansah & Ankoma-Sey, 2020). A lot of concerns have been raised about the validity and reliability of classroom observational data (Weisberg et al., 2009). In the USA, despite the proliferation of the use of these observation instruments in the school system, there is insufficient evidence of the reliability of their scores and other psychometric properties of such instruments (Patrick et al., 2020). In a similar vein, a query of the archives of some teacher training institutions in Ghana failed to find any evidence of the psychometric properties of their teaching practice rating scales (Oduro-Okyireh, 2020). Without the psychometric properties of measurement instruments, users of such instruments can neither decide on an instrument selection for teacher evaluation nor plan on an observation schedule. This implies that users may select and use observation instruments without knowing the consequences on learners and training programmes. This study emphasises the need for reliable observation data by investigating the reliability of the scores from an observation instrument (Intern Teaching Evaluation Form [ITEF]) used by a Ghanaian teacher training public university for its teaching practice programme (School Internship Programme [SIP]). The ITEF measures teachers' general competencies in instructional delivery.

This study fully used Generalisability Theory (GT), which is a psychometric model which is anchored on a statistical sampling technique which in a single analysis, partitions obtained scores into their primary sources of variation. GT provides a framework which is used to pinpoint and quantify the sources of measurement error based on which decisions can be taken to enhance the measurement procedure to give the most reliable scores (Li, 2022). Shavelson and Webb (1991), assert that, the Classical Test Theory (CTT) applies traditional approaches of estimating score reliability, and if applied, would estimate only a single source of measurement error at a time in a given analysis. This may be, differences in scores across occasions by the test-retest method, internal consistency of items on an assessment instrument by the split-half and Kuder-Richardson (KR 20 and KR 21) methods, or the consistency with which different raters score the same performance of students by Cohen's Kappa. Thus, the CTT in a single analysis, assesses only one type of consistency and then bulks all the different sources of error into a single undifferentiated measurement error.

It is due to this major limitation of CTT in the estimation of reliability that GT comes in. The conception and estimation of reliability by CTT is broadened by GT through the provision of a conceptual framework which is grounded in statistics, that allows a researcher to untangle numerous sources of error that make up the undistinguishable error (*E*) in CTT (Li, 2022; Webb & Shavelson, 2005). This study is an addition to the literature on the application of GT in measurement instrument development, as a resource material for researchers in the social sciences, especially in the Ghanaian educational system and other developing countries where GT application is not pervasive (Shavelson & Webb, 1991).

The study sought to use GT to investigate the score reliability of the ITEF which is used to evaluate the teaching practice performance of interns over repeated occasions. It was also to determine the

ISSN: 2408-7920

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

major source of error in the ITEF scores and an optimal scoring design for the ITEF. Since the inception of the use of the ITEF, these important psychometric indicators of the measurement procedure have not been investigated.

REVIEW OF RELATED LITERATURE

Conceptual framework of GT

According to Brennan (2001), CTT and Analysis of Variance (ANOVA) are from the parents of GT. The conceptual framework of GT is pinned on CTT and ANOVA and this gives birth to several conceptual and statistical issues which make up the framework (Brennan, 2001; Webb & Shavelson, 2005). According to Brennan (2001), GT is seen as an extension and advancement of CTT, which is made possible by applying measurement of variance techniques to measurement data. This is achieved by using factorial ANOVA to partition an individual's obtained score into effects for the true score, sources of error, and each of their combinations (Shavelson & Webb, 1991). The conceptual issues of the framework are classified into a universe of admissible observations and Generalisation (G) study, and a universe of generalisation and Decision (D) study, while the statistical issues are divided into variance components, error variance, and coefficients and indices.

Conceptual issues

Generalisability study and universe of admissible observations

By the assertion of Shavelson and Webb (1991), "from the perspective of GT, a measurement is a sample from a universe of admissible observations, observations that a decision maker is willing to treat as interchangeable to make a decision" (p.3). The universe in this context is taken as all the admissible conditions of a given facet of the measurement of interest. A facet is defined as a set of analogous conditions of measurement. For example, if an investigator wants to generalise from performance on a set of few occasions of teaching practice, to a larger set of occasions, which may be all the occasions of teaching in one's lifetime as a professional teacher, then, 'occasion' becomes a facet of the measurement procedure. Any set of occasions makes up an admissible condition of measurement for the occasion facet. The occasion universe will then be defined by the set of all admissible occasions. With this, the researcher's universe of admissible observations contains an infinite occasion facet. A generalisation (G) study is a study carried out to assess the variability of all the possible facets of a measurement procedure (Shavelson & Webb, 1991).

Universe of generalisation and decision study

The foremost aim of a G study is to give extensive information about the sources of variation in a given measurement and to obtain estimates of variance components associated with the universe of admissible observations. A G study should define the universe of admissible observations as largely as possible (Brennan, 2001; Webb & Shavelson, 2005).

A Decision (D) study adopts the information obtained from a G study on a measurement procedure to redesign the most efficient measurement procedure for a given purpose. A D study selects some desired facets for designated purposes, and by doing this, constricts the score interpretation to a

ISSN: 2408-7920

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

universe of generalisation, which is basically, the conditions of a facet to which a decision maker wants to generalise.

Statistical issues
Variance components and error variances
Variance components refer to the variances of effects in a G study after factorial ANOVA has been used to divide an individual's obtained score into effects for the facets and their combinations (Shavelson & Webb, 1991). Error variances, on the other hand, refer to the variance components of the various sources of variability in a given measurement apart from σ_p^2, which is the variance component for the universe score for the object of measurement, which is mostly humans in social science measurements.

Error variances result from the variations among persons in a group on a measured characteristic which are as a result of chance factors. Applying the appropriate GT analysis can estimate the variance components for all sources of error and then apportion them to the computation of two main kinds of error variances, which are absolute error variance and relative error variance (Brennan, 2001; Shavelson & Webb, 1991). If measurements are used to determine an individual's or group's absolute level of performance on a measured trait, it is referred to as absolute decisions. The variance of errors associated with these decisions are known as absolute error variance. On the other hand, when measurements are used to place individuals relative to each other based on performance on a measured trait, they give rise to relative decisions. The variance of errors associated with these types of decisions are known as relative error variance (Brennan, 2001; Shavelson & Webb, 1991).

Coefficients and indices
Three types of reliability-related coefficients are obtainable in GT. They are, the coefficient of relative measurement, coefficient of absolute measurement, and coefficient of criterion-referenced measurement (Cardinet et al., 2011). The coefficient of relative measurement refers to the proportion of total score variance that is accounted for by the true differences among randomly sampled objects of study. It gives the proportion of variability in individuals' observed scores that is systematic. Cardinet et al. (2011) assert that, this is the coefficient that Cronbach (1972) referred to as the G coefficient. It is denoted by $E\rho^2$ and it means the same as the coefficient of reliability in CTT. The G coefficient, $E\rho^2$, is useful when scores are to be given relative interpretations as in norm-referenced interpretations. It estimates how the measurement procedure can place the objects of measurement, which are mostly humans in social science research, relative to one another, and to estimate reliably the differences among them.

The coefficient of absolute measurement also called the dependability coefficient (D coefficient) (Brennan, 2001; Brennan & Kane, 1977), and denoted by Φ (Phi), assesses the capability of a measurement procedure to place the objects of measurement correctly on a scale in absolute terms (Cardinet et al., 2011). The D coefficient is useful in absolute interpretations, particularly important when it comes to domain-referenced decisions. The dependability index, Φ, is different from the generalisability coefficient, $E\rho^2$, because Φ uses absolute error variance, $\sigma^2(\Delta)$, while $E\rho^2$

ISSN: 2408-7920

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

uses relative error variance, $\sigma^2(\delta)$. In absolute decisions, the main effect of the task at hand, such as the difficulty level of a task, impacts absolute individual performance and so is critical in the delineation of measurement error. Since $\sigma^2(\Delta)$ is typically larger in value than $\sigma^2(\delta)$, the result is that Φ is usually smaller in value than $E\rho^2$ (Brennan, 2001; Marcoulides, 2000; Wiley et al., 2013).

The phi (lambda) coefficient, symbolised by $\Phi(\lambda)$, is a coefficient of criterion-referenced measurement and stretches further the Phi coefficient to encompass cut-off score uses in assessment. The $\Phi(\lambda)$ assesses the capability of a measurement instrument to reliably place individuals' scores in terms of an assigned standard of performance (cut-off score), which is set at λ on the measurement scale (Cardinet et al., 2011). For instance, in the teaching practice scores used for this study, the pass mark (cut-off score) is set at 50.0 points on a scale of 0 – 100 and hence, $\Phi(\lambda)$ becomes $\Phi(50)$ and it shows how dependably the ITEF could place individuals on one side of this point. It measures the interval between an individual score and the assigned cut-off score. It is, the dependability of the measured difference between an obtained score, x, and the cut-off score, S.

Empirical Review

Many empirical studies have been conducted worldwide on instrument development with the use of GT focusing on reliability investigation. Notable among them is the study by Patrick et al. (2020) with the aim of assessing the score stability of the Framework for Teaching (FFT). The FFT is an observation instrument used for teacher assessment in the USA. It comprises of four domains of practice, which 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 in the study. The stability of kindergarten teachers' classroom environment, instruction, and total FFT scores for reading and mathematics lessons were evaluated in the study. The scoring design was such that, each of the three scored 200 reading and mathematics lessons taught by 20 kindergarten teachers. The study adopted a two-facet (lessons, raters), partially nested (lessons within teachers), random design to partition the FFT's classroom environment, instruction, and total scores into probable sources of variability (teachers, lessons, raters, and their interactions).

The findings of the study were that, for reading and mathematics, the score variances accounted for by differences among teachers were 71% and 76% for the classroom environment, 49% and 37% for instruction, and 69% and 66% for the total scores respectively. On reliability estimates, G-coefficients had a range of 0.92 to 0.96 for the classroom environment and total scores, and 0.87 and 0.79 for reading and mathematics instruction respectively. D studies conducted found that two raters, each scoring three reading lessons or four mathematics lessons, are required to achieve satisfactory reliable total scores. For scores on instruction, three raters each scoring seven reading lessons are required, and finally, more than four raters each scoring eight lessons are required for mathematics to achieve satisfactory reliable total scores.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

Again, a study by Ramadhan et al. (2019), used GT to design a standard instrument for assessing physics teachers' competencies. The instrument is comprised of four main competencies which are pedagogical, personality competencies, social competencies, and professional competencies. The research subjects were 30 physics teachers and four experts as assessors. The study adopted a nested design for both the G study and the D-study. The G study adopted a two-facet p x (i:r) random effects model to compute variance components for a person, rater, item, person and rater interaction, and error.

The study found that the major source of error was the residual (52.3%) and the variance due to the object of measurement was 9.2%. The G coefficient for both relative and absolute measurements was 0.74. The D study conducted showed that to reach a G coefficient for relative interpretation of at least 0.70, which is appropriate for research applications, the evaluator 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 ought to have the design, P x (I: R), where I Random, R Fixed (P = 30, R = 4 and I = 4). Ramadhan et al. (2019) 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" (p. 336).

Further, Huijgen et al. (2017), undertook a similar study to craft a reliable observation instrument (Framework for Analysing the Teaching of Historical Contextualisation [FAT-HC]) and scoring rubric to assess the approaches by which history instructors teach historical contextualisation in their history classrooms. The study participants were five observers (raters), five teachers (object of measurement) and 265 students in the upper track of second cycle education `in Holland. According to Huijgen et al. (2017), "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" (p. 163). Each observer was made to rate two videotaped lessons of every teacher to give a total of 50 scores. Two designs were employed in the study. First, to investigate the dimensionality of the instrument, a univariate G-study at the item level which involved seven facets in a crossed design was performed. Second, to investigate the reliability of the instrument, a multivariate G study with a crossed (t × l × o), with history teachers (t), number of history lessons (l) and observers (o), was performed.

It was found that the item facet accounted for most of the variance (47.25 %) in the obtained scores indicating that the instrument is unidimensional when used to evaluate how history teachers teach historical contextualisation in their history classrooms. Again, the teacher facet accounted for most of the variability (59.1%) in the obtained scores which showed a high reliability of the FAT-HC. The other percentages of total variance were 34.7%, 4.58% and 1.63% for the residual, observers and lessons respectively. A D-study indicated that the ideal scoring design would use two observers who would each rate two lessons taught by the same teacher ($\Phi = 0.83$) or three observers who would each rate the same lesson taught by one teacher ($\Phi = 0.80$).

ISSN: 2408-7920

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

METHODOLOGY

Research design

The paradigm of positivism was adopted for the study. Positivism explains reality, as anything that can be perceived with the human senses, not depending on human consciousness, is ordered and is objective (Khan & Mohsin Reza, 2022). The methodology chosen for the study is quantitative. This option was dictated by the nature of the data (scores) to be collected, the data collection instrument (rating scale which quantifies traits), and the nature of data analysis (statistical analysis). The above processes form the research methodology for the study.

The GT design adopted for this study was a random effects one-facet crossed design. Interns (*p*) taught on three occasions (*o*) and each intern was rated by one mentor (rater, *r*). The rater facet has one level and this does not meet a basic assumption for the application of GT analysis. Different lessons were taught on different occasions by different interns with no uniformity and hence the lesson (l) facet could not fit into an appropriate GT design. The 25 items on the ITEF for a sample size of 9082 for three occasions would have given 681,150 item scores to make it unbearable to analyse. Hence, the rater, lesson and item facets could not be included in the analysis. They were considered as unmeasured facets in the study. The design of the study eventually was an intern (*p*) crossed with the occasion (*o*), and symbolically denoted by (*p*×*o)*.

Given the design, an obtained score for an intern on an occasion, $X_{po,}$ is decomposed into the following four effects:

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

The obtained score equation for a one-facet crossed design can be written by grouping the terms as follows:

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

In this design, an effect, aside from the grand mean, has a distribution (Marcoulides, 2000; Shavelson & Webb, 1991). Each distribution has a mean of zero and variance σ^2. The total variance of a collection of obtained scores, X_{po}, of all persons and occasions in the universe is given by the summation of the three variance components: $\sigma^2(X_{po}) = \sigma^2_p + \sigma^2_o + \sigma^2_{po,e.}$ This shows that the variance of occasion scores in a one-facet fully crossed design can be divided into three sources of variation as a result of differences between persons, occasions and the residual.

Population

The target population for the study was all regular bachelor's degree graduates from the 46 departments in the 14 faculties of the university up to the 2017/2018 academic year. The accessible population was all regular bachelor's degree graduates from 46 departments in the 14 faculties,

ISSN: 2408-7920

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

for three academic years from 2015/2016 to 2017/2018, who were 18,339 in number. The regular bachelor's degree graduates were used for the study because, unlike the university's sandwich and distance students, the regular students participate in an officially planned internship programme as part of their undergraduate programme. The internship programme covers one academic semester and interns are rated by school mentors on three occasions. We selected three years for the study because of the intention to examine the psychometric properties of the ITEF scores over a time frame and three years was deemed adequate.

Sample and sampling procedure

A total of 9,082 bachelors' degree graduates from eight purposively selected academic faculties for the 2015/2016 to 2017/2018 academic years were selected by the census method. The teaching practice triplicate scores of the 9082 bachelor's degree graduates were selected for the study. The 9,082 triplicate scores resulted in 27,246 individual scores for three occasions of rating. Each intern was rated on three occasions with a total of 9,082 observers in the study (one observer rated one intern). We selected the 2015/2016 academic year as the beginning point for data collection because, by the end of this academic year, only eight faculties of the university had graduated final year students and had students' internship scores that could be accessed. The teaching subject areas of students in these eight faculties are archetypal of all the academic courses that the university offers for teacher training.

Selection and training of observers (raters)

Observers (raters) in the study were purposively selected from partnership schools nationwide where students undertake their internship programme. They were selected based on possession of a minimum bachelor's degree in teacher education, with a specialised subject content area. They were then given a one-day intensive workshop on the use of the ITEF in teaching practice observation. The training session is done annually for new observers and as a refresher for already trained ones. At the start of the internship period, each intern is assigned one observer (mentor) in the school of his/her choice, who supervises every aspect of the intern's work in the school and rates his/her teaching on three occasions for formal evaluation purposes. In addition, one trained university observer (supervisor) visits every school once for monitoring and evaluation purposes which include carrying out a lesson observation session with the school observer to authenticate the scores assigned by the mentor to the intern.

Data collection instrument

Existing records of teaching practice scores of bachelor's degree graduates selected for the study were collected and used for data analysis. The scores were obtained from the use of an already existing observation instrument (ITEF). The ITEF is divided into five sections and each has some sub-elements that are scored on a five-point scale ranging from zero (0) to four (4). A score of zero (0) indicates the non-existence of the skill while four (4) indicates the ideal skill. The first section is on "Planning and Preparation" and has a maximum score of 12. This section centres on lesson planning and preparation with selection of appropriate teaching and learning materials (TLM's) for teaching. This stage comes before practical teaching begins. There are three sub-elements in this section. The second section deals with "Instructional Skills" with a maximum score of 40. At

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

this stage, practical teaching begins in the classroom and the rater starts to rate the intern's lesson as it unfolds. There are ten sub-elements in this section that the rater should observe attentively to allot scores. The third section addresses "Classroom Management" with a maximum score of 16. It centres on the ideal rapport that should exist between the teacher and the students and how the teacher uses this rapport to manage and control the classroom during instructional delivery. There are four sub-elements in this section that the rater must observe attentively to allot scores.

The fourth section is on "Communication Skills" with a maximum score of 16. This addresses the ideal communication style between the teacher and the students by which knowledge is imparted. Hence, it requires paying rapt attention to both the written and spoken communication of the teacher in order to evaluate their correctness. This section has four sub-elements. The last section is "Evaluation" with a maximum score of 16. This section has four sub-elements. It requires that the rater pays attention to the lesson closure and all the evaluation strategies of the teacher in order to allot scores. The total score for rating of each lesson using the ITEF is 100.

Data collection procedure

Data were obtained from the use of the ITEF observation instrument administered over three occasions. In all, 9,082 triplicate teaching practice scores for the academic years 2015/2016 to 2017/2018, giving rise to 27,246 scores, were obtained from the teaching practice unit of the university for analysis. Data collection lasted for a period of one month.

Data processing and analysis

Data analysis was done by performing a univariate generalisability analysis using EduG statistical programme (Cardinet et al., 2011). The faculties were represented by their respective thematic course areas or specialties. G coefficients for relative ($E\rho^2$) and absolute (Φ) interpretations and variance components, with their standard error of measurement (SEM) were computed. This is because the focus of the study was on both the stability of the scores and absolute performance of the interns across occasions. The G coefficient for relative ($E\rho^2$) interpretation was to enable a description of the percentage of observed score variance accounted for by systematic variations in interns' knowledge of subject matter and skills in teaching and also give an index of the quality of the measurement design on a scale of 0.0 - 1.0 (Marcoulides, 2000). Further, the G coefficient for absolute (Φ) interpretation was to enable a description of the dependability of the ITEF scores (Brennan, 2001).

RESULTS AND DISCUSSION

Reliability of the ITEF scores for each academic year from 2015/2016 to 2017/2018

ISSN: 2408-7920

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

The first specific objective of the study was to find the extent of reliability (stability and dependability) of the ITEF scores across the three occasions of rating for each academic year from 2015/2016 to 2017/2018. Table 1 shows the G coefficients (relative and absolute) for the ITEF scores for each faculty and academic year from 2015/2016 to 2017/2018.

Table 1: G coefficients for ITEF scores from 2015/2016 to 2017/2018 academic years

	Academic Year					
	2015/2016		2016/2017		2017/2018	
Faculty/ Specialty	$E\rho^2$	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

From Table 1, the G coefficient, $E\rho^2$, 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. For 2017/2018, it is from 0.67 for social science to 0.81 for natural science. Considering Phi coefficient, Φ, for 2015/2016, the range is from 0.59 for technical to 0.78 for natural science. For 2016/2017, the range is from 0.66 for applied science to 0.81 for natural science, foreign languages and vocational. For 2017/2018, it is from 0.67 for social science to 0.81 for natural science.

It is seen that, for each academic year from 2015/2016 to 2017/2018, for relative interpretation, the ITEF scores are strongly reliable for the fact that the G coefficients range from 0.66 to 0.84. For absolute interpretation, the ITEF scores are moderately dependable for technical and strongly dependable for all other thematic course areas (0.59 – 0.78) for the 2015/2016 academic year, while strongly dependable (0.67 – 0.81) for all other thematic course areas for 2016/2017 and 2017/2018 academic years.

As stated by Shavelson and Webb (1991), the Coef_G relative, $(E\rho^2)$, of 0.66 to 0.84 represents the proportion of observed score variance due to systematic differences in interns' knowledge of subject content and proficiency in teaching. It is the interns' universe-score variability. It also gives an index of the quality of the measurement procedure by use of the ITEF on a scale of 0.0 - 1.0 (Marcoulides, 2000). The Coef_G absolute (Φ) of 0.59 – 0.81 is an index of the dependability of the ITEF scores obtained from the measurement design. It represents the accuracy of generalising from an intern's obtained score on a single occasion to the average score that the intern would have received under all possible occasions of lesson delivery (Marcoulides, 2000; Shavelson & Webb,

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

1991). Hence, on a scale of 0.0 - 1.0, it could be concluded that the ITEF scores are dependable to a great extent.

The purpose, methodology by way of GT application and the first finding of this study are in line with those of a study by Patrick et al. (2020), which evaluated the score stability of the Framework for Teaching (FFT). G coefficients for relative interpretation ($E\rho^2$) had a range of 0.92 to 0.96 for the FFT's classroom environment and total scores, and were 0.87 and 0.79 for reading and mathematics instruction correspondingly. These indices indicate higher quality (score stability) of the FFT scores just as indicated by the results of the current study with the ITEF. The G coefficients reported with the FFT are relatively higher than that of the ITEF. The obvious reason is the differences in the contexts of assessment in the two studies. Again, unlike the current study, Patrick et al. (2020) was only interested in the FFT's score stability and so did not report on G coefficients for absolute interpretation.

The set of first findings (stability and dependability) of the current study is again consistent with that of Ramadhan et al. (2019), who applied GT to craft a standard instrument for assessing physics teachers' competencies. G coefficients for both relative and absolute interpretations were 0.74. Some values of G coefficients recorded by the current study are larger while others are smaller than the recorded value of Ramadhan et al. (2019). Again, the differences in the assessment contexts and especially, the sample sizes used for the two studies account for the differences in the G coefficients. Ramadhan et al. (2019) used a partially nested two facet p x (i:r) random effects model with a sample size of only 30 teachers, while the current study used a random effects one-facet crossed design with sample sizes for the eight faculties ranging from 61 to 1274. Smaller sample sizes have debilitating effects on G and Phi coefficients, as indicated by Atilgan (2013) that, G and Phi coefficients computed for a sample size of 30 was less than the G and Phi parameters, with relative root mean square error (R-RMSE) value greater than 0.01.

Major source of error in the ITEF scores for each academic year from 2015/2016 to 2017/2018

The second specific objective of the study was to find the major source of error in the ITEF scores. This was achieved by examining the estimated variance components and standard error of measurement (SEM) of the occasion facet and the residual to find the main source of error in the ITEF scores for each academic year from 2015/2016 to 2017/2018. The SEM is extremely important when evaluating the measurement precision of an instrument (Cardinet et al., 2011). In GT analysis, it is the SEM that shows the magnitude of the error affecting the results in terms of relative and absolute interpretations (Cardinet et al., 2011; Cronbach, 1972). Table 2 shows the estimated variance components for an intern (p), occasion (o) and residual (po,e) for each academic year and speciality with their proportions of total variance and SEMs.

Table 2: Estimated variance components for intern (p), occasion (o) and residual (po,e)

Estimated Variance Components

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

Speciality / Proportion / Standard Error	2015/2016			2016/2017			2017/2018		
	Intern (p)	Occasion (o)	Residual (po,e)	Intern (p)	Occasion (o)	Residual (po,e)	Intern (p)	Occasion (o)	Residual (po,e)
Applied Science	27.19	61.96	28.38	32.96	0.52	50.50	25.82	1.09	30.43
Proportion	48.8	0.2	51.0	39.2	0.6	60.1	45.0	1.9	53.1
Standard Error	3.346	0.178	1.806	4.111	0.485	2.873	2.801	0.835	1.645
Business	15.6	5.34	14.24	20.88	3.40	14.73	16.73	1.94	18.17
Proportion	44.2	15.6	40.2	53.5	8.7	37.7	45.4	5.3	49.3
Standard Error	1.032	3.929	0.504	1.462	2.423	0.585	1.412	1.392	0.783
English and Communication	17.63	14.35	15.94	19.68	1.98	22.33	14.13	2.00	13.54
Proportion	36.8	29.9	33.3	44.7	4.5	50.8	47.6	6.7	45.6
Standard Error	3.123	1.025	1.520	4.964	1.662	2.860	2.178	1.479	1.106
Foreign Languages	31.02	0.15	31.44	34.15	0.21	24.22	31.58	0.18	28.55
Proportion	49.5	0.2	50.2	58.3	0.4	41.3	52.4	0.3	47.3
Standard Error	5.585	0.307	2.958	3.253	0.197	1.310	3.130	0.186	1.519
Natural Science	28.94	0.07	24.00	34.79	0.91	23.35	36.69	0.65	15.78
Proportion	54.6	0.1	45.3	58.9	1.5	39.5	58.1	1.0	40.8
Standard Error	3.728	0.133	1.697	3.648	0.704	1.406	3.172	0.504	1.267
Social Science	18.61	0.61	27.27	19.35	0.27	26.63	20.03	0.19	29.05
Proportion	40.0	1.3	58.6	41.8	0.6	57.6	40.7	0.4	59.0
Standard Error	1.673	0.467	1.135	1.415	0.213	0.922	1.208	0.152	0.814
Technical	12.77	9.18	17.16	25.10	5.46	23.30	13.72	3.94	12.73
Proportion	32.6	23.5	43.9	46.6	10.1	43.3	45.1	13.0	41.9
Standard Error	1.233	6.515	0.792	2.127	3.893	1.052	1.553	2.819	0.769
Vocational	32.76	5.81	32.67	23.69	3.05	13.46	15.55	2.95	17.03
Proportion	46.0	8.2	45.9	58.9	7.6	33.5	43.8	8.3	47.9
Standard Error	7.009	4.404	3.676	3.824	2.241	1.289	3.276	2.226	1.837

From Table 2, the estimated variance components for interns (p) (estimated universe score variances), give the actual variations among the interns (p) in terms of differential levels of knowledge of subject matter and pedagogical skills and hence this variance does not constitute an error. The estimated variance component for the occasion (o) represents measurement error as a result of differences in occasional conditions of instructional delivery and evaluation of teaching. The estimated variance component for the residual represents measurement error as a result of the p x o interaction and unidentified sources. Thus, the two error variances are the variance components for the occasion (o) and the residual (po,e).

Table 2 shows that, in all the academic years, the variance component (error variance) for the residual (po,e) is larger for all the thematic course areas. In the 2015/2016 academic year, it ranges from 15.94 (33.3% of total variance) for English and communication to 27.27 (58.6% of total variance) for social science. In the 2016/2017 academic year, it ranges from 13.46 (33.5% of total

ISSN: 2408-7920

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

variance) for applied science to 50.50 (60.1% of total variance) for vocational. In 2017/2018 academic year, it ranges from 15.78 (40.8% of total variance) for natural science to 29.05 (59.0% of total variance) for social science.

Considering the SEM associated with the estimated variance components, Table 2 shows that, there are five cases where the occasion (o) facet has the largest SEM values. These are, for 2015/2016, for business and technical, SEM of 3.929 and 6.515 respectively. For 2016/2017, for business and technical, SEM of 2.423 and 3.893 respectively. For 2017/2018, for technical, SEM of 2.819. Further, residual (po,e) has the largest SEM values in 19 out of the 24 cases of G study analysis, indicating a greater error margin accounted for by the residual. It could therefore, be concluded that, for each academic year from 2015/2016 to 2017/2018, the major source of error in the ITEF scores is the p x o interaction and unidentified sources (residual). This is followed closely by the occasion facet in each academic year.

Different occasions would bring diverse instructional times, durations and conditions of lesson delivery and assessment. These would result in generalisation from the occasion sample to the occasion universe being less accurate (Webb et al., 2006). In the 2015/2016 academic year alone, this error of generalisation from the occasion facet is given as 0.1% to 29.9%. Again, error variance from the p x o interaction is explained as the erraticism that results from variations from one occasion to another in a particular intern's behaviour (Brennan, 2001). Consequently, the pairing of an intern's eccentricities with a given occasion creates an interaction between interns and occasions which gives rise to inconsistencies from one occasion to another in a particular intern's behaviour. This upsurges variability and makes generalisation from an intern's score on an occasion to his average score over all possible occasions in the occasion universe less accurate.

In this study, the percentage of error from the p x o interaction is relatively large. For instance, it is 33.5% to 60.1% for the 2016/2017 academic year. The obvious explanation is that this study is a one-facet one which used only the occasion facet with any other facets such as the rater, item and lesson, taken as unmeasured facets. Undoubtedly, these are sources of non-sampled systematic and random vacillations at play in the analysis that swell up the error margin. Brennan (2001) and Cardinet et al. (2011) assert that, in a one-facet crossed design such as this, after the error due to the occasion facet has been accounted for, it cannot be detected whether further variations in occasion scores represent the p x o interaction or random unknown sources of variation. Therefore, these two sources of variability are combined as a residual and delineated by the $p \times o$ interaction clouded by other sources of variability.

The finding above is also in line with that of Huijgen et al. (2017) who aimed at developing the FAT-HC observation instrument. In their study, the major source of error was the residual (34.7%), just as it is with the current study. This finding is also in line with that of Ramadhan et al. (2019), who found that the major source of error in their study to develop an instrument for assessing physics teachers' competencies was the residual (52.3%).

An optimum number of occasions of rating is needed to obtain the most reliable ITEF scores and ensure economy in the use of resources in the SIP

ISSN: 2408-7920

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

The third specific objective of the study was to find the optimal number of occasions of rating needed to obtain the most reliable (stable and dependable) ITEF scores. This was achieved by running a D study (optimisation) with the results of the G study, by varying the number of occasions of rating of teaching practice from 1 to 6, to give new values of G coefficients ($E\rho^{2*}$ and Φ^{*}), for the combined scores of all three academic years for all eight faculties. The asterisk (*) indicates D study coefficients. The relative and absolute SEM values were also computed to help explain the margins of error associated with the D coefficients.

The rate of occurrence of G coefficients of at least 0.80 for a given number of occasions was used as an index for judging an optimum number of occasions (Burns, 1998; Cardinet et al., 2011). A particular number of occasions was chosen as optimum for a given academic year if more than half (i.e., more than four) of the eight faculties reached Coef_G relative and absolute of at least 0.80 at that number of occasions. By the assertion of Brennan and Kane (1977) and Webb et al. (2006), G coefficients should be at least 0.80 to be satisfactory for formative evaluations. A G coefficient of at least 0.80 was used as the standard because the teaching practice exercise is of a formative kind with the intent of improving teaching skills. The D study results from the combined scores of the three academic years are shown in Table 3.

Table 3: G coefficients of D study for numbers of occasions for ITEF scores from 2015/2016 to 2017/2018 academic years for all faculties

No of Occasion	**1**		**2**		**3**		**4**		**5**		**6**	
G Coef.	$\mathbf{E\rho^{2*}}$	$\mathbf{\Phi^{*}}$	$\mathbf{E\rho^{2*}}$	$\mathbf{\Phi^{*}}$	$\mathbf{E\rho^{2*}}$	$\mathbf{\Phi^{*}}$	$\mathbf{E\rho^{2*}}$	$\mathbf{\Phi^{*}}$	$\mathbf{E\rho^{2*}}$	$\mathbf{\Phi^{*}}$	$\mathbf{E\rho^{2*}}$	$\mathbf{\Phi^{*}}$
	.52	.51	.68	.68	.77	.76	.81	.81	.84	.84	.87	.86
Rel./ Abs. SEM	5.24	5.32	3.71	3.76	3.03	3.07	2.62	2.66	2.35	2.38	2.14	2.17

From Table 3, for four occasions of rating, both Coef_G relative ($E\rho^{2*}$) and Coef_G absolute (Φ^{*}) attained the threshold of 0.80 (i.e., 0.81 each). It could also be seen that as the number of occasions increases, both the relative and absolute SEM decrease indicating increased measurement precision as a result of a decrease in error in the measurement procedure. For example, for four occasions, relative SEM and absolute SEM are 2.62 and 2.66 respectively, which are less than the values for three occasions which are 3.03 and 3.07 respectively. These results indicate that the levels of precision for both relative and absolute measurements of the ITEF are much less reliable on fewer occasions. Five and six occasions have lesser values of both relative and absolute SEM, but since for four occasions, the acceptable reliability criterion of at least 0.80 has been achieved, and we want to economise the use of resources to reduce cost in the SIP, four occasions of rating would be accepted. Four occasions of rating are therefore, taken as the optimum number of occasions for the most stable and dependable ITEF scores.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

The finding above is the ultimate finding that lays the basis for the enhancement of the measurement procedure of the ITEF. The approach, purpose and the third finding of the current study are consistent with those of Ramadhan et al. (2019), whose D study indicated that, to reach a G coefficient for relative interpretation ($E\rho^2$) of at least 0.70, being acceptable for research purposes (Brennan & Kane, 1977), the assessor should use four items. It must be noted that this study was on the development of a new instrument and so did not have any prior scoring design. Again, unlike the current study, the study of Ramadhan et al. (2019) was for research purposes and therefore, accepted the G coefficient for relative interpretation ($E\rho^2$) of at least 0.70 as appropriate.

Again, the approach, purpose and the third finding of the current study are in line with those of a study by Patrick et al. (2020). With their Framework for Teaching (FFT), D studies indicated that two raters, each scoring three reading lessons or four mathematics lessons, are required to obtain satisfactorily reliable total scores. For scores on instruction, three raters each scoring seven reading lessons are required, and finally, more than four raters each scoring eight lessons are required for mathematics to achieve satisfactorily reliable total scores. This scoring design was recommended by Patrick et al. (2020) in place of the original three observers each scoring 200 reading and mathematics lessons delivered by 20 KG teachers for the classroom environment and instruction domains of the FFT.

Finally, the approach, purpose and the third finding of the current study are in line with those of a study by Huijgen et al. (2017), in the development of the FAT-HC observation instrument. A D-study indicated that the optimum scoring design would involve two observers with each evaluating two lessons taught by the same teacher ($\Phi = 0.83$) or three observers with each evaluating the same lesson taught by one teacher ($\Phi = 0.80$). This scoring design was recommended in place of the original scoring design where each rater rated two lessons of each teacher to give a total of 50 ratings.

CONCLUSION

The aim of the study was to use Generalisability Theory (GT) to investigate the reliability (stability and dependability) of the ITEF scores. A total of 9,082 triplicate bachelor's degree teaching practice scores for three academic years were used for the analysis. Notwithstanding the treatment of certain facets (i.e., rater, lesson and item) of the measurement procedure as unmeasured facets in the analysis, which come as a limitation in the study, the ITEF scores were found to be strongly stable and moderately to strongly dependable, for three occasions of rating for three academic years from 2015/2016 to 2017/2018. The establishment of these psychometric properties of the ITEF is formidable enough to guarantee trust and confidence in the use of the scores by decision-makers. For the fact that the ITEF scores are moderately to strongly dependable, employers can use this index for employment decisions for students who obtained their bachelor's degrees in the academic years used in the study. The developers of the ITEF should formally document these important psychometric properties of the scores in reference to the ITEF used for the evaluation

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

of teaching practice. This will give an insight into the quality of the scores and the measurement design to increase trust in their use, and to help in improvement of the teaching practice programme.

Again, the finding that the major source of error in the ITEF scores for the three occasions of rating for the academic years from 2015/2016 to 2017/2018, is the p x o interaction followed by the occasion facet for each academic year, is an evidential depiction of the potency of GT over CTT in reliability estimation. It has laid bare a source of error in the ITEF scores, taking cognisance of the limitations of this study. This lays the foundation for further study on the measurement procedure of the ITEF in the SIP.

Finally, for the three occasions of rating for the academic years 2015/2016 to 2017/2018, the ideal number of occasions of rating of teaching practice for the most reliable (stable and dependable) ITEF scores was found to be four. It is based on empirical evidence that four occasions will achieve the acceptable G coefficients required for formative evaluation and since this is the minimum number of occasions for achieving these acceptable G coefficients, it would help economise the use of resources in the SIP. This justifiably calls for the improvement of the measurement procedure which employs the use of the ITEF in the SIP. It is, therefore, recommended that the implementors of the teaching practice programme should increase the number of occasions of evaluation from three to four. This is a clarion call for more commitment on the part of raters (mentors) as this will indubitably increase their workload.

Limitations of The Study

Only the occasion facet was used in the analysis of the current study. The rater, lesson and item facets were excluded from the G study analysis. This resulted in the estimated variance component of the intern (p) by occasion (o) interaction (residual) confounded with presumed errors from other unmeasured facets in the study. The analysis model used, therefore, could not fully untangle the errors of measurement and spread them to as many sources as possible to result in a comparatively smaller variance from the residual. Consequently, a conclusion on one major source of error, being a single facet in the entire ITEF scores could not be made.

Practical Implications

Teachers who receive thorough and reliable evaluations are better equipped to adopt effective teaching practices. This has a positive impact on student learning and engagement. The development of competent and confident teachers contributes to the long-term success of the education system. Students benefit from higher-quality education, which can improve their academic and social outcomes.

Social Implications

ISSN: 2408-7920

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

An optimally designed evaluation form can help reduce biases related to race, gender, and socioeconomic status. This promotes a more inclusive educational environment where all student teachers are given equal opportunities to succeed. Teachers trained through equitable evaluation practices are more likely to be sensitive to the needs of diverse student populations. This can lead to more inclusive and supportive classroom environments. Reliable evaluation data can inform policy decisions related to teacher certification, professional development, and educational standards. Policymakers can use this data to develop evidence-based policies that support teacher quality and student achievement.

Further Research

To have a holistic insight into the main source of error in the ITEF scores, it is recommended that this study is replicated using at least two facets including the occasion (o), item (i) and the rater (r).

REFERENCES

Atilgan, H. (2013). Sample Size for Estimation of G and Phi Coefficients in Generalizability Theory. *Eurasian Journal of Educational Research (EJER)*(51).

Brennan, R. L. (2001). *Generalizability Theory*. Springer Science & Business Media.

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

Burns, K. J. (1998). Classical reliability: using generalizability theory to assess dependability. *Research in nursing & health, 21*(1), 83-90.

Cardinet, J., Johnson, S., & Pini, G. (2011). *Applying generalizability theory using EduG*. Routledge.

Cronbach, L. J. (1972). The dependability of behavioral measurements. *Theory of generalizability for scores and profiles*, 1-33.

Huijgen, T., van de Grift, W., Van Boxtel, C., & Holthuis, P. (2017). Teaching historical contextualization: the construction of a reliable observation instrument. *European Journal of Psychology of Education, 32*(2), 159-181.

Khan, K. K., & Mohsin Reza, M. (2022). Social Research: Definitions, Types, Nature, and Characteristics. In *Principles of Social Research Methodology* (pp. 29-41). Springer.

Li, G. (2022). How many students and items are optimal for teaching level evaluation of college teachers? Evidence from generalizability theory and Lagrange multiplier. *Sustainability, 15*(1), 2.

Marcoulides, G. A. (2000). Generalizability theory. In *Handbook of applied multivariate statistics and mathematical modeling* (pp. 527-551). Elsevier.

Oduro-Okyireh, G. (2020). Dependability of Students 'Internship Mentors 'Results Using Generalizability Theory at University of Education, Winneba.

Patrick, H., French, B. F., & Mantzicopoulos, P. (2020). The reliability of framework for teaching scores in kindergarten. *Journal of Psychoeducational Assessment, 38*(7), 831-845.

ISSN: 2408-7920

African Journal of Applied Research
Vol. 10, No. 1 (2024), pp. 57-74
http://www.ajaronline.com
http://doi.org/10.26437/ajar.30.06.2024.04

Quansah, F., & Ankoma-Sey, V. R. (2020). Evaluation of pre-service education programme in terms of educational assessment. *Int. J. Res. Teach. Educ*, *11*, 56-69.

Ramadhan, S., Nasran, S. A., Utomo, H. B., Musyadad, F., & Ishak, S. (2019). The implementation of generalisability theory on physics teachers' competency assessment instruments development. *International Journal of Scientific and Technology Research, 8*(7), 333-337.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. SAGE.

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: overview. *Encyclopedia of statistics in behavioral science*.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). 4 reliability coefficients and generalizability theory. *Handbook of statistics*, *26*, 81-124.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.

Wiley, E. W., Webb, N. M., & Shavelson, R. J. (2013). The generalizability of test scores. In *APA handbook of testing and assessment in psychology, Vol. 1: Test theory and testing and assessment in industrial and organizational psychology.* (pp. 43-60). American Psychological Association.

ISSN: 2408-7920