British Journal of Education
Vol.12, Issue 7, 115-135, 2024
Online ISSN: 2054-636X (Online)
Print ISSN: 2054-6351(Print)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development-UK

Evaluation of Teaching Practice: Assessment of the Dimensionality of the Intern Teaching Evaluation Form (ITEF)

George Oduro-Okyireh
Department of Interdisciplinary Studies, Akenten Appiah-Menka University of Skills Training and Entrepreneurial Development (AAMUSTED), Ghana
Email: gookyireh@aamusted.edu.gh

doi: https://doi.org/10.37745/bje.2013/vol12n7115135 Published June 15, 2024

Citation: Oduro-Okyireh G. (2024) Evaluation of Teaching Practice: Assessment of the Dimensionality of the Intern Teaching Evaluation Form (ITEF), *British Journal of Education*, Vol.12, Issue 7, 115-135

ABSTRACT: *This study aimed at investigating the dimensionality of the Intern Teaching Evaluation Form (ITEF), which is an observation instrument used in the evaluation of teaching practice at a public university in Ghana. The instrument was used to assess the teaching skills of a group of interns across three lessons. Data were obtained over a three-year period from 2016/2017 to 2018/2019 academic years during the Students Internship Programme (SIP) of the university. Part of the overall aim was to find the score stability (reliability) of the ITEF when it is used to evaluate teaching skill. A random effect two facets fully crossed design where intern (t) was crossed with item (i) and lesson (l), given by (t×i×l) was adopted for the dimensionality investigation. Further, a random effect one-facet fully crossed design, where inter (t) was crossed with lesson (l), given by (t×l) was adopted for the reliability investigation. The sample sizes for the study were 9132 undergraduate bachelor's degree ITEF triplicate scores for the reliability investigation and 50 ITEF triplicate scores for the dimensionality investigation, selected by purposive and simple random sampling methods respectively. Univariate generalisability analysis using EduG was performed to analyse data. Findings were that, the ITEF is not unidimensional, but highly reliable (strongly stable) across the delivery of three lessons. Caution should be exercised in taking the ITEF scores as unidimensional in running correlational analyses with other criteria, unless this is done only from the perspective of 'use,' which should be one of the important considerations in instrument design.*

KEYWORDS: Teaching practice, Dimensionality, Reliability

INTRODUCTION

Since 1835 that the first teacher training college was established by the Basel Mission in Ghana (Micots, 2015), observation instruments have been used in the evaluation of teaching competencies in the Ghanaian education system. With the generic teaching

competency requirements which the teacher training curriculum places on teacher trainees, attention has always been placed on the development of instruments which are aimed at the evaluation of teachers' generic teaching competencies (Fortin & Legault, 2014). In Ghana, different teacher training institutions have their own observation instruments, such as the Intern Teaching Evaluation Form (ITEF) used by the University of Education, Winneba (UEW), the Internship Record Book (IRB) used by the Akenten Appiah-Menka University of Skills Training and Entrepreneurial Development (AAMUSTED), and the Supported Teaching in School Assessment Form (STSAF) used by the Colleges of Education (CoEs) in Ghana (Oduro-Okyireh, 2020).

According to Brion and Cordeiro (2019), in the USA, there are several observation instruments that teacher training institutions use to evaluate the acquisition of teaching skills by upcoming teachers admitted into teacher education programmes. These instruments aim at competence in scheme of work preparation, lesson notes preparation, conduciveness of the classroom environment to instruction, and instructional methods used to facilitate students' learning. Some of these instruments are the Candidate Preservice Assessment of Student Teaching (CPAST), the Educative Teacher Performance Assessment (edTPA), and the Resident Educator Summative Assessment (RESA), just to mention a few.

Although, the crafting and use of classroom observation instruments could be extremely labour-intensive, they are undoubtedly more impartial in data collection to examine classroom teacher behaviour because they give accounts of events as they naturally unfold during instructional sessions (Pianta & Hamre, 2009; Wragg, 2011). The crafting and usage of classroom observation instruments are indispensable in directing teacher education programmes and evaluating interventions in the classroom (Darling-Hammond, 2012; Good & Lavigne, 2015).

The indispensability of classroom observation data in teacher training is corroborated by Anastasi and Urbina (1997) that, "direct observations of behaviour play an essential part in personality appraisal, whether in the clinic, counselling centre, classroom, personnel office, or any other context calling for individual evaluations" (p. 463). According to Oduro-Okyireh (2020), evaluation of teacher trainees by raters adopting direct observation is acceptable by both theoretical assessment principles and practical convenience because of the nature of the trait being measured. It is not practically expedient to assess pedagogical skills with achievement tests because all such tests are evaluations of minute behaviour samples put up in artificial settings at given times. Such obtained behaviour samples must be evaluated under standardised conditions to be sure of their validity and for the assignment of appropriate interpretations before they can be practically meaningful. Direct observational methods for evaluation of behaviour (as

used in direct observation of teaching practice) however, provide a comprehensive sampling of behaviour which occurs in natural settings.

A lot of uncertainties have been raised concerning the structural validity and reliability of classroom observational data (Galvão et al., 2021; Hill et al., 2012). Patrick et al. (2020) assert that, in the USA, in spite of the explosion of the use of observation instruments in the school system nationwide, there is insufficient evidence of the validity and reliability of the instruments, including other psychometric properties of such scales which are so important to aid decision making. Oduro-Okyireh (2020) emphasises that, this is apparently not different from the Ghanaian situation, since the field of psychometry is relatively an emerging area in psychology and education in Ghana.

This study addresses the need for structurally valid observation instruments for observational studies by investigating the dimensionality of an observation instrument (Intern Teaching Evaluation Form [ITEF]) used by a Ghanaian teacher training public university for its teacher training internship programme. The ITEF is supposedly designed to measure teachers' competence in instructional delivery. Dimension operationally refers to the attributes that researchers think play a major role in the data they obtain from study subjects (Irribarra & Arneson, 2023). Does the ITEF actually measure the dimension of the trait—teachers' competence in instructional delivery?

Structural validity is the extent to which the scores of a given measurement instrument are a true and satisfactory depiction of the dimensionality of the construct being measured (Pradhan & Hati, 2022). Assessment of a quantitative instrument's structural validity is mainly done by collecting and analysing data using statistical tests and other measures to assess the accuracy with which the instrument measures the dimension which it is purported to measure (Reichenheim et al., 2014).

Based on the recommendations of Hill et al. (2012), I applied Generalisability Theory (GT) to investigate the dimensionality and reliability of the ITEF (Cronbach, 1972; Shavelson & Webb, 1991). GT is a psychometric theory which is founded on a statistical sampling method, which in a one-shot analysis, has the ability to divide observed scores into their primary sources of variation. It must be noted that, the conception and estimation of reliability by classical test theory (CTT) is widened by GT through the modelling of a conceptual framework, founded in statistics, that allows an investigator to untangle multiple error sources that constitute the undistinguishable error (E) in CTT (Brennan, 2001; Li, 2022; Shavelson & Webb, 1991). Primarily, GT offers a framework that is used to detect and estimate the sources of errors of measurement in a measurement procedure, after which decisions can be made to improve it so as to give more reliable scores and at the same time economise the use of resources (Li et al., 2014).

Statement of the Problem

From the onset of the use of the ITEF observational instrument for evaluation of teaching practice, there has not been any literature on its dimensionality and reliability. Does the ITEF measure a single dimension of a single construct called teaching skill and as such

be classified as unidimensional (Van der Lans et al., 2018) or multiple dimensions of multiple skills and so be described as multidimensional? Again, when the ITEF is used to evaluate teaching skills of interns on teaching practice by one rater across different lessons, there would undoubtedly be measurement inconsistencies. In a summary, and expressed in question form and divided into two sub-problems, the problem of the study is: What is the dimensionality of the ITEF? How reliable is the ITEF when it is used to evaluate teaching practice? GT provides a formidable framework for answering these questions and so it formed the main conceptual base on which this study was pinned.

Research Questions

The main purpose of this study was to investigate the dimensionality of the ITEF with its reliability as an inherent aim. To achieve this central aim, I raised the following research questions to guide the study:

i. What is the dimensionality of the ITEF when it is used to rate teaching skill?
ii. What is the extent of reliability of the ITEF when it is used to rate teaching skill?

Assumptions of the Study

As theorised by Hoyt and Melby (1999) and Brennan (2001), the elementary assumptions underlying the use of GT to do G study analysis are:

i. data used in G studies should be either interval or ordinal. This study used measures of teaching skill which are interval;
ii. obtained scores are made up of universe scores and at least a source of error. In this study, errors are assumed to stem from the item and lesson facets;
iii. errors of measurement are assumed to be independent of the universe score and uncorrelated among themselves. All effects in the measurement model are independent of each other;
iv. samples used in estimation of the variance components, are random samples from their individual populations. However, fixed facets can be used;
v. the measured attribute is in a steady state, and any differences in obtained scores on different occasions are due to at least one source of error, and not due to systematic variations in the person due to maturation. Data in this study were taken at closer time intervals to curtail score variability due to maturation.

THEORETICAL FRAMEWORK

Dimensionality of Measurement Instruments

Interest in the issue of dimensionality of measurement instruments and measurement data arose probably due to the advancement in the work on item response theory (IRT), which was initially developed in the early 1950s by Frederic Lord and other psychometricians (Luecht & Hambleton, 2021). Most IRT models assume that the construct fundamental to test performance is unidimensional (He & Min, 2024).

What is the concept of dimensionality as applied to measurement in psychology and education? Earlier literature on the concept of dimensionality has largely centred on the harmonisation between theoretical and statistical perspectives. In one perspective,

investigators understudy the theoretical attributes conjectured to underpin assessment results. In another perspective, investigators understudy the dimensional structure as a rudimentary assumption underlying the statistical models used to generate results. When dimensionality is used in these senses, it functions interchangeably with the concept of psychological or educational construct (Irribarra & Arneson, 2023).

Earlier, Reckase (1990) had advanced an argument in support of the view above and added that, the terminologies, "unidimensional" and "multidimensional" have often not been used with the precision that they need. This ultimately has resulted in a lot of confusion. He continues that, there are two common uses of the term dimensionality when it is used in reference to psychological and educational assessments. “First, dimensionality is used to refer to the number of conjectured psychological constructs that are believed to be required for successful performance on a test” (p. 3). An example is that, numerical calculation and verbal reasoning are needed to do well on mathematical word problems. In this case, numerical calculation and verbal reasoning are seen as two psychological dimensions which are theorised to exist to give a sound explanation to variations in performance on the assessment task. This use of dimensionality is referred to by Reckase (1990) as ‘psychological dimensionality.’ Second, dimensionality is used to refer to the smallest number of mathematical variables that is required to summarise a matrix of item response data. For example, a vector which has two elements may be needed in a probabilistic model of test performance to accurately predict how an individual will respond to a particular set of assessment tasks. This use of the term dimensionality is referred to as ‘statistical dimensionality’(Reckase, 1990).

Irribarra and Arneson (2023) advance the argument further that, all educational assessments are often subject to restrictions imposed by time, resources, legal requirements, and also the users’ ability to understand the results and use them appropriately. These restrictions function in isolation of the theoretical specifications or the perfect fit of statistical models. They therefore, argue that a third viewpoint of ‘use’ is needed to address the complexities presented by dimensionality. Hence, integrating the perspectives of ‘theory,’ ‘statistics’ and ‘use,’ is crucial to comprehend thoroughly the complications that surround dimensionality. These three viewpoints, then present an expressive justification that reflects the restraints imposed by diverse groups who ideally should partake in the processes of developing an assessment instrument. These groups should include researchers, psychometricians and users. Consistency between these strands is important in matching the controlling principles for the design of the assessment, the statistical models used to interpret results, and the practical application of those results. In coordinating the three perspectives in instrument development, Irribarra and Arneson (2023) conclude that, instrument designers ought to make the effort to harmonise these perspectives of dimensionality during the design stage. Theoretical dimensionality enables practical interpretations, providing the conceptual basis for inference. Statistical dimensionality provides an experiential foundation establishing the methodological support for inferences. Finally, utilitarian dimensionality results in usability, spelling out an application context’s practical requirements.

Psychological and educational constructs need to be clearly conceptualised before they be measured accurately. One fundamental requirement in conceptualising constructs is to determine whether they are unidimensional or multidimensional (Irribarra & Arneson, 2023; Trumpp et al., 2015). A unidimensional construct is the one that has a single underlying dimension of measurement and so can be measured with one instrument. Examples of unidimensional constructs are, subtraction of two-digit numerals, identification of place values of digits in numerals, and reading comprehension. A multidimensional construct on the other hand, has two or more underlying dimensions of measurement. For example, if a person's mathematical aptitude is conceptualised as consisting of four dimensions – addition, subtraction, multiplication and division, then mathematical aptitude is termed as a multidimensional construct. In this case, each of the underpinning dimensions needs to be measured distinctly, using different tests or a single test with distinct parts measuring each dimension, for the four scores to be combined to give a composite score for the mathematical aptitude construct (Irribarra & Arneson, 2023; Trumpp et al., 2015).

The argument above is corroborated by Brown et al. (2023) and Durvasula et al. (2006) that, the justification for unidimensionality is that, the meaning of any measure is clearest if only a single dimension underpins it. They continue that when a measure addresses only a single dimension, the correlation between that measure and a criterion is clearer. On the other hand, when more than one dimension is present, which connotes the assessment of more than one variable, the correlation between that measure and a criterion may be confounded. Therefore, the establishment of the dimensionality of an assessment instrument is a necessity for internal consistency, construct validity, and theory testing.

Technically speaking, moving beyond psychological dimensionality to statistical dimensionality, the supposition of unidimensionality is not easy to confirm (Van der Lans et al., 2018). A multiplicity of tests have been proposed to evaluate unidimensionality, but there is no agreement on any single statistical approach (Timmerman et al., 2018). There is hence, an extensive ongoing deliberation on the best criterion to evaluate the goodness of fit of statistical models. A school of thought recommends the use of exact tests, which can reject the null-hypothesis of unidimensionality. Another school of thought argues that an exact test of unidimensionality is too stringent and frequently rejects the null-hypothesis even when the given data lend themselves to an appropriate description of one dimension. They therefore, recommend the use of 'approximate fit' indexes, of which they propose the use of 'root mean square error of approximation' (RMSEA) or a method which uses some type of ratio between eigenvalues (Van der Lans et al., 2018).

Further, on ascertaining dimensionality of measurement instruments, a well proven method is GT, which recognises that multiple sources—persons, raters, lessons and dimensions, in the case of measurement with the ITEF, can contribute to measurement error (Cronbach, 1972; Shavelson & Webb, 1991). This enables the assessment of the

combined and interaction effects of these sources across different occasions. Facets in a measurement procedure in GT are classified into two types: facets of generalisation and facets of differentiation (Durvasula et al., 2006). Facets of generalisation contribute to measurement error resulting in error variances. In the use of the ITEF, raters and lessons constitute facets of generalisation. It is important that, the measurement procedure is designed to reduce variance emanating from these facets. Facets of differentiation on the other hand, represent the set of objects whose attribute is being measured and who are to be compared in the study. In the case of the measurement with the ITEF in this study, the facets of differentiation are the set of interns and the set of items on the ITEF.

As not all persons are alike, their responses to external prompts across lessons are bound to differ. Hence, a differentiation facet such as the set of interns contribute to variance that is expected. This variance actually shows the true differences between the interns in terms of differences in abilities in teaching (Oduro-Okyireh, 2020). It is important that measurement instruments are crafted to maximise variance from the facet of differentiation (Durvasula et al., 2006). The variance component which accrues from the facet of differentiation is known as the universe score variance (Shavelson & Webb, 1991). Since the set of items on the ITEF is being tested for its dimensionality, it is also considered as a facet of differentiation in this study.

Further, on usage of statistical tests to ascertain dimensionality, the position of Shavelson and Webb (1991), (Brennan, 2001) and Huijgen et al. (2017) is clear. That is, in estimation of variance components in GT analysis, if an assessment instrument is unidimensional and it is used in a given measurement procedure, then the items on the instrument (as a facet of differentiation), should account for the greater part of the total variance. Other facets in the measurement procedure (including the interaction effects) should account for a lesser part of the total variance.

Lastly, considering the issue of 'use' as a perspective in instrument development, it must be understood that, multidimensionality is complicated, due to the fact that it depends on the purpose for which the assessment instrument was designed (Linacre, 2023). In exemplification, an arithmetic test (addition, subtraction, multiplication, division) is said to be unidimensional from the point of view of the school administrator who is deciding whether a pupil should be promoted to the next grade-level. However, the same test is multidimensional from the perspective of the school psychologist who is diagnosing specific learning difficulties in the pupil. In conclusion, Van der Lans et al. (2018) assert that, observations of teaching behaviours in teacher development can be ascribed to a single latent construct or dimension, and that is, teaching skill.

Reliability of Observation Instruments

There are complexities involved in measuring a construct in social science research, and therefore, it is not satisfactory just to measure a construct using any scale that is desired. These scales must be tested to be certain that they undoubtedly measure the unobservable construct that was purported to be measured (i.e., the scales are 'valid'), and also, they measure the purported construct consistently and exactly (i.e., the scales are 'reliable')

(Rwegoshora, 2016). Reliability and validity, are the basic psychometric characteristics of all measurement scales, and are the indices by which the precision and accuracy of measurement procedures are assessed in scientific research.

It was in 1904 that Spearman advanced a logical and mathematical argument that all test scores are 'imperfect measures' of human traits and that, the observed correlation coefficient between such imperfect test scores is an underestimation of the correlation between 'true objective scores.' In explaining the terms 'imperfect measures' and 'true objective scores,' Spearman put forward the idea of correction of correlation coefficient for attenuation because of less reliable measurement instruments (Crocker & Algina, 1986). Again, he put forward the method for obtaining the coefficient of reliability needed in making such correction. According to Crocker and Algina (1986), the work of Spearman in principle, brought forth the classical test theory (CTT), on which the classical true score model is pinned.

The main idea of the CTT model that was advanced by Spearman was that a given obtained score could be seen as the addition of two hypothetical scores (a component of a true score and a random error score) (Crocker & Algina, 1986; Finch & French, 2018). Mathematically, this is given as: $X = T + E$, where X is the obtained score, T is the individual's true score, and E is a random error component. Hence, CTT divides the obtained score variance, σ_X^2, into two parts which are the true score variance, σ_T^2, and random error score variance, σ_E^2.

In CTT, the concept of reliability is defined fundamentally as the ratio of the true score variance to the obtained score variance. Different methods of reliability estimators have been put forward by psychometricians based on this rudimentary definition. The concept of reliability can be divided into three broad groups. These are stability, internal consistency, and inter-rater reliability. Stability and internal consistency estimators are based directly on classical definition of reliability while inter-rater reliability represents a modern measure of reliability (Finch & French, 2018). Stability estimators are the test-retest method, coefficient of equivalence, and alternative forms method. These methods aim at assessing the capability of a measurement instrument to give consistent results when conducted at different times with similar conditions of administration. A measurement instrument is said to be reliable if it gives similar results, as assessed by an appropriate computed correlation coefficient. Stability estimators assess measurement error as a result of inconsistency in forms (equivalence) and time (test-retest). The main problem with these methods is the difficulty of guaranteeing similar conditions of instrument administration and an appropriate time interval between consecutive administrations (Crocker & Algina, 1986; Webb et al., 2006).

Internal consistency estimators are the split-half method and its correction to double length (Surhone et al., 2010), Kuder-Richardson formula (Kuder & Richardson, 1937), coefficient alpha (Cronbach, 1951) and omega (McDonald, 2013). These reliability estimators assess the degree to which a set of items on an instrument are consistent with

a designated task when the instrument is administered on a single occasion. These estimators assess measurement error emanating from inconsistency in sampling the item domain. The main problem is that these estimators thrive on the supposition that the items are continuous or interval. This is not always the case. To curtail the problem imposed by this assumption, Zumbo et al. (2007) assert that latest developments in measurement theory have resulted in ordinal versions of alpha and omega.

Inter-rater reliability estimators evaluate the degree to which given scores from two or more raters observing the same activity using the same scoring rubric are in agreement with each other. They evaluate measurement error emanating from inconsistencies among raters. Finch and French (2018) assert that, the percentage of agreement among raters, Cohen's Kappa for two raters (Stemler, 2007), and multiple-rater Kappa (Fleiss, 1981) have been used primarily to assess interrater reliability. It is worthy of noting, that these estimators are instructively logical, but are theoretically dissimilar from the well-known definition of reliability and should not be taken as alternatives for reliability estimates in describing observational instruments (Crocker & Algina, 1986; Lamm et al., 2020).

The first two categories of reliability estimators above, are only appropriate for relative decisions and interpretations. Again, they can be used to assess only one kind of measurement error in a given analysis, with each type of estimate determining the extent to which true scores deviate from obtained scores. According to (Brennan, 2001), Shavelson and Webb (1991), and Suen and Lei (2007), the problem is that unlike GT, CTT does not possess the ability to evaluate inconsistencies in test forms, items, raters, administrators, dimensions or occasions in a single analysis.

For all measurement designs in GT, a general reliability index called generalisability coefficient (G coefficient) can be calculated, which is given as in CTT by the ratio of estimated true score variance to estimated total obtained score variance. Its value ranges from 0.0 to 1.0 and higher values closer to 1.0 are indications of more reliable measurement procedures (Cardinet et al., 2011; Marcoulides, 2000). Shavelson and Webb (1991) and Cardinet et al. (2011) call the attention of all, to the fact that, the definition of the G coefficient in GT is subject to the intended purpose of a given measurement. This is because error variance differs from relative to absolute decisions.

According to Cardinet et al. (2011), three types of reliability-related coefficients are obtainable in GT. They are, the coefficient of relative measurement, the coefficient of absolute measurement, and the coefficient of criterion-referenced measurement. The G coefficient of relative measurement computes the percentage of total score variance which emanates from the true differences between randomly sampled members of a differentiation facet. It reflects the percentage of variance in individuals' obtained scores which is systematic. This coefficient is given by $E\rho^2$ and it means the same as the coefficient of reliability in CTT. The G coefficient, $E\rho^2$, is used if scores are to be given relative interpretations as in the case of norm-referenced interpretations. It gives an

estimate of how accurately the measurement procedure is able to put the members of the differentiation facet in the order of performance and to estimate reliably the differences between them.

The G coefficient of absolute measurement is defined as the dependability coefficient (D coefficient) (Brennan, 2001; Brennan & Kane, 1977), and given in symbol by Φ (Phi). It is used to locate the members of a differentiation facet reliably on a scale in absolute terms (Cardinet et al., 2011). This coefficient is applied when making domain referenced decisions. It uses all estimated variance components in its calculation with the exception of that of the object of measurement.

The dependability coefficient, Φ, is different in interpretation from the generalisability coefficient, $E\rho^2$, in the sense that Φ uses absolute error variance, $\sigma^2(\Delta)$, while $E\rho^2$ uses relative error variance $\sigma^2(\delta)$ in their computations. With absolute decisions, the main effect of the trait being measured, for instance, difficulty level of an item, influences total performance of individuals and hence this is significant in the definition of measurement error. For the tact that $\sigma^2(\Delta)$ is generally larger than $\sigma^2(\delta)$, the consequent effect is that Φ is usually lesser in value than $E\rho^2$ (Brennan, 2001).

The Phi(lambda) coefficient, Φ(λ), is a coefficient of criterion-referenced measurement. It extends the G coefficient of absolute measurement (Phi coefficient [Φ]) to include cut-off score applications (Cardinet et al., 2011). According to Cardinet et al. (2011), the Φ(λ) indicates how reliably a measurement instrument locates an individual's score with respect to a cut-off score which is set at λ on the measurement scale. For instance, with the ITEF, the pass mark (cut-off score) is 50.0 points on a 0 – 100 scale. Therefore, Φ(50) indicates how dependably the measurement procedure places individual interns on one side of this point. It gives an estimate of the interval from the individual scores to the selected cut-off score. Fundamentally, it is the dependability of the measured interval between an obtained score x and the cut-off score S.

METHODOLOGY

Research Design

For the dimensionality investigation, a random effect two-facets fully crossed design was adopted. Interns (*t*) taught three lessons (*l*) and each intern was rated by one rater (*r*) using all 25 items (*i*) on the ITEF. The G study analysis was done at the item level using the item (*i*) as the differentiation facet. Thus, the design was, item *(i)* crossed with intern (*t*) and lesson *(l)*. It is given in symbols by $(i \times t \times l)$.

In this design, an observed score for one item, intern, and lesson, $X_{itl,}$ can be decomposed into seven different effects. Each effect, apart from the grand mean has a distribution with mean zero and variance component σ^2 (Marcoulides, 2000; Shavelson & Webb, 1991). The total variance of a distribution of obtained scores, Xit_l, over all items, interns, and lessons in the universe is given by the sum of the seven variance components:

$$\sigma^2(X_{itl}) = \sigma^2_i + \sigma^2_t + \sigma^2_l + \sigma^2_{it} + \sigma^2_{tl} + \sigma^2_{il} + \sigma^2_{itl,e}$$

Thus, the variance of item scores in a two-facet fully crossed design can be partitioned into seven sources of variation due to differences in items, interns, lessons, their interactions and the residual (item by intern by lesson interaction combined with unidentified error sources).

For the reliability investigation, a random effect one-facet fully crossed design was adopted. Interns (t) taught three lessons (l) and rated by one rater (r) using all 25 items on the ITEF. For every lesson of each intern, an overall score for the 25 items on the ITEF was computed and used as a composite score for the analysis. This is because using each score of the 25 items for the 9,132 interns for three lessons each, would have given 684,900 scores, which would have been unbearable to analyse. Hence, the design used was, interns (t) crossed with lesson (l), and symbolised by ($t \times l$).

In this design, each effect, apart from the grand mean has a distribution with mean zero and variance component σ^2 (Marcoulides, 2000; Shavelson & Webb, 1991). The total variance of a distribution of obtained lesson scores, X_{tl}, over all interns and lessons in the universe is given by the sum of the three variance components:

$$\sigma^2(X_{tl}) = \sigma^2_t + \sigma^2_l + \sigma^2_{tl,e}$$

Hence, the variance of lesson scores in a one-facet fully crossed design can be divided into three sources of variation stemming from differences in interns, lessons and the residual. It must be noted that, in the two designs for this study, the rater facet had only one level and this violates a rudimentary assumption of GT analysis and so was excluded from the analysis. Thus, the rater facet was handled as an unmeasured facet in the entire study.

Sample and Data Collection

Purposive sampling method was used to select eight out of 14 academic faculties of the university and three academic years from 2016/2017 to 2018/2019. From the eight faculties for the three academic years, the census method was then used to select 9132 undergraduate bachelor's degree ITEF triplicate scores for the reliability investigation of the ITEF. Further, simple random sampling method was used to select 50 out of the 9132 ITEF triplicate scores for the dimensionality investigation of the ITEF. I selected the 2016/2017 academic year as the starting point for data collection because permission for data collection was granted in 2019 and an inherent aim of the research was to study the psychometric properties of the ITEF scores over a three-year period, and hence, 2016/2017 academic year as a starting point was deemed appropriate. Only eight faculties were selected for the study because their teaching subject areas are representative of all the academic courses that the university offers for teacher training. Adding more faculties would have been a duplication of specialties. Table 1 gives the distribution of students for the three academic years by the eight faculties as at the end of 2018/2019 academic year.

British Journal of Education
Vol.12, Issue 7, 115-135, 2024
Online ISSN: 2054-636X (Online)
Print ISSN: 2054-6351(Print)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development-UK

Table 1
Distribution of students by academic year and faculty

Faculty	Year			
	2016/ 2017	2017/ 2018	2018/ 2019	Total
1. Agriculture Science	309	342	257	908
2. Business Education	633	538	808	1979
3. Education and Communication Sciences	61	150	117	328
4. Foreign Languages and Communication	342	353	119	814
5. Social Science	835	1274	582	2691
6. Science and Environment	276	414	201	891
7. Technical Education	490	274	472	1236
8. Vocational Education	109	86	90	285
Total	3055	3431	2646	9132

From Table 1, the sample sizes for the faculties for the academic years range from 61 for the Faculty of Education and Communication Sciences, for 2016/2017 academic year to 1274 for the Faculty of Social Science, for 2017/2018 academic year. The sample sizes have a range of 1213.

Selection and Training of Observers (Raters)
Observers (raters) in the study were purposively selected from partnership schools countrywide where students have chosen to undertake their internship programme. Eligibility criteria were the possession of a minimum of bachelor's degree in teacher education and a specialised subject content area. They were then given a day's intensive workshop on the use of the ITEF in observing teaching practice. The workshop is organised yearly for new observers and as a refresher for already trained ones. At the beginning of the internship period, each intern is assigned one observer (mentor) in the school, who supervises every aspect of the intern's work in the school and rates their teaching on three occasions for formal evaluation purposes. In addition to this, a trained university observer (supervisor) visits every school once during the internship session for monitoring and evaluation purposes which include carrying out a lesson observation session with the school observer for authentication of the scores given to the intern.

Structure of the Observation Instrument
The principles for the development of the ITEF were the optimum requisite skills and generic professional competencies required to be exhibited by a teacher during a teaching session. The ITEF has five sub-sections, in which each section has a number of sub-elements that are rated on a five-point scale which ranges from zero (0) to four (4). Section One is on "Planning and Preparation" with maximum points of 12. This section addresses lesson planning and preparation with selection of appropriate teaching and learning materials (TLM's) for a lesson. It comes before practical classroom instructional delivery. The section contains three indicators which are: Exhibits knowledge of subject matter; Objectives are "SMART" and align instructional strategies with lesson

objectives; and Content connects with and challenges students' present knowledge, skills and values.

Section Two addresses "Instructional Skills" with maximum points of 40. Hands-on instructional delivery starts at this stage in the classroom and the rater starts to assess the intern's teaching skill as it unfolds. The section contains ten indicators that the rater should pay critical attention to them in order to award scores. These are: States purpose, objectives, and procedures for lessons; Gives procedural and instructional directions clearly; Uses a range of strategies for whole class, small group and individual teaching/learning; Motivates students; Relates lesson to prior knowledge and life experience; Presents lesson in a systematic manner; Uses effective questioning techniques of the level of students; Engages students in critical thinking and problem solving; Uses techniques that modify and extend student learning; and Engages students in lesson closure.

Section Three addresses "Classroom Management" with maximum points of 16. It centres on the appropriateness of the rapport that ought to exist between the teacher and the students and how the teacher uses this rapport to manage the classroom during instructional delivery. The section contains four indicators that the rater should pay critical attention to them in order to award scores. These are: Manages classroom routines effectively; Respects diversity among students; Maintains positive rapport with students; and Knows each student as an individual.

Section Four addresses "Communication Skills" with maximum points of 16. This centres on teacher-student interaction by which impartation of knowledge occurs. The section contains four indicators. These are: Communicates with confidence and enthusiasm; Communicates at students' level of understanding; Uses accurate non–verbal, oral/sign and written communication; and Projects voice/hand shapes/orientation appropriately.

Section Five addresses "Evaluation" with maximum points of 16. It is required that the rater focuses on both the informal and formal, formative and summative evaluation approaches of the intern, which should start right from the commencement to the conclusion of instruction. The rater's duty is to align the instructional objectives to the evaluation strategies and items in the lesson and what the teacher does both during and after teaching. The rater must find out whether each instructional objective is fully evaluated. The section contains four indicators. These are: Monitors student's participation and progress; Provides immediate and constructive feedback; Bases evaluation on instructional goals/objectives; and Uses formal/informal assessment strategies to assess student learning before/during/after instruction to enhance learning. The total ITEF score for a lesson is 100%.

Data Processing and Analysis

British Journal of Education
Vol.12, Issue 7, 115-135, 2024
Online ISSN: 2054-636X (Online)
Print ISSN: 2054-6351(Print)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development-UK

The EduG statistical programme (Cardinet et al., 2011) was used in performing a univariate generalisability analysis. Estimation of variance components was done for the differentiation facets (person and item) and lesson facet, together with their interactions.

RESULTS

Research question 1

What is the dimensionality of the ITEF scale when it is used to rate teaching skill?

Research question 1 sought to find whether the ITEF actually measures the dimension of the one construct, teaching skill, which it has been designed to measure, by determining whether it is unidimensional. Table 2 shows the variance decomposition for item (*i*), intern (*t*), lesson (*l*) and residual (*itl,e),* with item *(i),* as a differentiation facet.

Table 2

Variance decomposition for Item (i), teacher (t), lesson (l) and residual (itl,e)

				Components				
Source	SS	Df	MS	Random	Mixed	Corrected	%	SE
Item (i)	64.1269	24	2.6720	0.0149	0.0149	0.0149	5.4	0.0050
Intern (t)	15.5669	49	0.3177	0.0005	0.0005	0.0005	0.2	0.0010
Lesson (l)	0.1365	2	0.0683	-0.0002	-0.0002	-0.0002	0.0	0.0001
i*t	348.6464	1176	0.2965	0.0199	0.0199	0.0199	7.3	0.0047
i*l	18.3435	48	0.3822	0.0029	0.0029	0.0029	1.1	0.0015
t*l	21.6235	98	0.2206	-0.0006	-0.0006	-0.0006	0.0	0.0013
i*t*l	556.5632	2352	0.2366	0.2366	0.2366	0.2366	86.1	0.0069
Total	1025.007	3749					100%	

Generalizability coefficients:
Coef_G relative: 0.83
Coef_G absolute: 0.83

Table 2 shows the G-study variance decomposition using 3750 scores obtained from 50 interns who taught three lessons each and rated with the 25-item ITEF scale. The item facet used as the differentiation facet explains only 5.4% of the total variance, while the item by intern by lesson explains as much as 86.1% of the total variance. The intern facet explains only 2% of the total variance.

Following the assertions of Shavelson and Webb (1991), Brennan (2001) and Huijgen et al. (2017), if the instrument is unidimensional, the differentiation facet (item), should account for the greater part of the total variance and the other facets (including the interaction effects) should account for a lesser part of the variance. As shown in Table 1, the item (*i*) accounts for only 5.4% of the total variance in the entire analysis, indicating that the ITEF scale is not one-dimensional with respect to its usage in rating teaching skill during teaching practice in the teacher preparation programme. It can therefore be concluded that, the ITEF is not unidimensional.

British Journal of Education
Vol.12, Issue 7, 115-135, 2024
Online ISSN: 2054-636X (Online)
Print ISSN: 2054-6351(Print)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development-UK

Research question 2

What is the extent of reliability of the ITEF when it is used to rate teaching practice?

Research question 2 sought to find out how reliable the ITEF is when it is used to rate teaching skill. With intern (*t*) as the differentiation facet, 27396 scores obtained from 9132 interns from eight faculties for three academic years, who taught three lessons each, were analysed in a G study. Table 3 shows the variance decomposition for the intern (*t*) and lesson (*l*) and their interactions (with the proportion of total variance in parenthesis for each facet) in the study.

Table 3

Variance decomposition for intern (t), lesson (l) and residual (tl,e)

Faculty / Specialty	Estimated Variance Components (Proportion of Total Variance [%])								
	2016/2017			2017/2018			2018/2019		
	Intern (t)	Lesson (l)	Residual (tl,e)	Intern (t)	Lesson (l)	Residual (tl,e)	Intern (t)	Lesson (l)	Residual (tl,e)
Applied Science	30.96 (36.9)	2.52 (3.0)	50.50 (60.1)	25.82 (45.0)	2.09 (3.6)	29.43 (51.3)	62.96 (53.6)	28.19 (24.0)	26.38 (22.4)
Business	21.88 (56.1)	1.40 (3.6)	15.73 (40.3)	17.73 (48.1)	0.94 (2.6)	18.17 (49.3)	19.64 (55.8)	3.34 (9.5)	12.24 (34.8)
English and Communication	19.66 (44.7)	2.98 (6.8)	21.35 (48.5)	16.13 (54.3)	2.00 (6.7)	11.54 (38.9)	20.63 (43.1)	11.35 (23.7)	15.94 (33.3)
Foreign Languages	37.15 (63.4)	1.21 (2.1)	20.22 (34.5)	32.58 (54.0)	1.18 (2.0)	26.55 (44.0)	33.02 (52.7)	2.15 (3.4)	27.44 (43.8)
Natural Science	35.79 (60.1)	1.91 (3.2)	21.35 (36.2)	36.69 (69.1)	1.65 (3.1)	14.78 (27.8)	34.94 (65.9)	2.07 (3.9)	16.00 (30.2)
Social Science	20.35 (44.0)	0.27 (0.6)	25.63 (55.4)	21.03 (42.7)	0.19 (0.4)	28.05 (56.9)	16.61 (35.7)	1.61 (3.5)	28.27 (60.8)
Technical	26.10 (48.5)	4.46 (8.3)	23.30 (43.3)	15.72 (51.7)	3.94 (13.0)	10.73 (35.3)	17.16 (43.9)	9.18 (23.5)	12.77 (32.6)
Vocational	22.69 (56.4)	5.05 (12.6)	12.43 (33.5)	14.55 (40.9)	3.97 (11.2)	17.04 (47.9)	34.76 (48.8)	4.81 (6.8)	31.67 (44.46)
Generalisability coefficients:									
Coef_G relative: 0.67 – 0.84									
Coef_G absolute: 0.67 – 0.81									

A reliable measurement instrument should have a higher percentage of the variance accounted for by differences in the observed teachers and a low percentage of the variance accounted for by lessons and observers (Brennan, 2001; Hill et al., 2012; Huijgen et al., 2017; Smit et al., 2017). From Table 3, taking each academic year into perspective, for 2016/2017, the estimated variance components for the differentiation facet (intern, t) range from 30.96 (36.9% of total variance) for Applied Science to 37.15 (63.4% of total variance) for Foreign Languages. A comparison of the estimated variance components for the two facets and their residual (tl,e) reveals that in five of the eight faculties, which are Business, Foreign Languages, Natural Science, Technical and

Vocational, the estimated variance components for the differentiation facet form the larger proportions of the total variances.

For 2017/2018, the estimated variance components for the differentiation facet (intern, t) range from 14.55 (40.9% of total variance) for Vocational to 36.69 (69.1% of total variance) for Natural Science. A comparison of the estimated variance components for the two facets and their interaction reveals that, in four of the eight faculties, which are English and Communication, Foreign Languages, Natural Science and Technical, the estimated variance components for the differentiation facet form the larger proportions of the total variances. For 2018/2019, the estimated variance components for the differentiation facet (intern, t) range from 20.63 (43.1% of total variance) for English and Communication to 34.94 (65.9% of total variance) for Natural Science. A comparison of the estimated variance components for the two facets and their interaction reveals that, in seven of the eight faculties, which are Applied Science, Business, English and Communication, Foreign Languages, Natural Science, and Vocational, the estimated variance components for the differentiation facet form the larger proportions of the total variances.

In total, out of 24 G study analyses in Table 3, 16 of them have the estimated variance components of the differentiation facet (intern, *t*), forming the larger proportions of the total variances. None of them has the estimated variance component of the lesson (*l*), forming the larger percentage of the total variance, and eight of them have the estimated variance components of the residual, (tl,e), forming the larger percentages of the total variances. In the 16 cases, the differences between the observed teachers (intern, *t*) accounted for 43.1% to 69.1% of the total variances.

It can therefore, be concluded that the ITEF is reliable since the observed interns (*t*) explain the largest variances in the largest number of cases of analyses than the lesson (*l*) and their interaction (Brennan, 2001; Hill et al., 2012; Huijgen et al., 2017; Smit et al., 2017). This is also supported by the G – coefficients reported at the bottom of Table 3 which show Coef_G relative: 0.67 – 0.84, and Coef_G absolute: 0.67 – 0.81. Table 2 also shows G coefficients for both relative and absolute interpretations as 0.83. These coefficients are indications of higher reliability of the ITEF.

DISCUSSION

This study aimed at establishing the dimensionality of the ITEF and also find the extent of its reliability when it is used to evaluate teaching practice. GT analysis was applied to provide indicators that the instrument is not unidimensional when it is used to evaluate teaching practice. A lesser percentage of the total variance (5.4%) was accounted for by the item facet (Brennan, 2001; Huijgen et al., 2017). GT analysis also indicated that a large percentage of the instrument's variance was accounted for by the differences in the observed interns while a smaller percentage of the variance was accounted for by the differences in lessons and the intern by lesson interaction, which is an evidence that the instrument is reliable (Brennan, 2001; Hill et al., 2012; Huijgen et al., 2017; Smit et al., 2017). Computed G coefficients indicate higher reliability of the ITEF.

The first finding suggests that in terms of psychological dimensionality, the ITEF has a number of conjectured psychological constructs that competence in them are required for a successful performance in a session of teaching practice (Reckase, 1990). Therefore, in conceptualising the educational construct measured by the ITEF by way of determining its dimensionality so that it can be measured accurately (Irribarra & Arneson, 2023; Trumpp et al., 2015), it could be concluded from this finding that the ITEF is not unidimensional. This is given credence by the composition and nature of the ITEF itself. This instrument is composed of five sections with each section having a number of items under it. The sections which are: planning and preparation; instructional skills; classroom management; communication skills; and evaluation, which are combined to assess competence in instructional delivery, have been found statistically in this study as not measuring the dimension of a single construct labelled teaching skill, but supposedly the dimensions of different educational constructs.

From this finding, it is seen that, the argument advanced by Brown et al. (2023) and Durvasula et al. (2006) in favour of unidimensionality, that, when a measure addresses only one dimension, its correlation with a criterion is clearer, does not go in favour of the ITEF. Hence, users of the ITEF scores should treat each section as a measure of the dimension of a separate construct for the purpose of ascertaining internal consistency, construct validity, and theory testing. However, considering the issue of 'use,' as a perspective in instrument development, which is stringently tied to the purpose for which instruments are designed (Linacre, 2023), the ITEF can be taken as unidimensional. This is by following the assertion of Van der Lans et al. (2018) that, the different sections of the ITEF which require demonstration of skills in them as part of teaching behaviours in teacher development, can be ascribed to a single latent dimension, which is, teaching skill. Caution should be exercised however, in taking ITEF scores as unidimensional when finding correlations with other criteria, unless it is done only from the perspective of 'use.'

The second finding of this study is that, the ITEF is highly reliable. Substantiating this finding, Brennan and Kane (1977) assert that, G coefficients must be at least 0.70 for research purposes, at least 0.80 for formative evaluations, and at least 0.90 for summative evaluations. Webb et al. (2006) add that, G coefficients of at least 0.80 are considered satisfactorily reliable to make decisions on individuals with respect to their obtained scores, even though a higher value of 0.90, is ideal if the decisions have momentous ramifications. It is evidently clear that the acceptable minimum reliability threshold of at least 0.70 (0.67 corrected to one decimal place, with many faculties obtaining at least 0.80)) has been obtained to give credence to the use of the ITEF for academic purposes.

According to Cardinet et al. (2011), the value of $0.67 \leq E\rho^2 \leq 0.84$ represents the proportion of variance in individuals' obtained scores which is systematic, and indicates the degree to which the measurement procedure that uses the ITEF is able to distinguish reliably between the members of the differentiation facet (interns) in the study. It is an estimate of how accurately the measurement procedure can locate the interns, in the order

of their relative teaching skills, and to evaluate correctly the intervals between them. Marcoulides (2000) asserts that it also offers a practical coefficient of the quality of the measurement design using the ITEF on a scale of 0.0 - 1.0. Hence, the conclusion can be drawn that the measurement design which uses the ITEF is high in quality.

The G coefficient for absolute interpretation, which is the dependability coefficient, $0.67 \leq \Phi \leq 0.81$, on the other hand, locates the interns dependably on a scale in absolute terms (Cardinet et al., 2011). Because this is applied in absolute decisions, the main effect of the attribute being measured, which in this case is the ability of the interns to meet the requirement of each item on the ITEF during teaching, is key in determining performance of individuals and so this plays a formidable role in the characterisation of measurement error. This dependability coefficient also reflects the accuracy of generalising from an intern's obtained score in one lesson to the average score the intern would have received under all the possible lesson deliveries in his/her professional life as a teacher, taking the universe of generalisation to be infinite (Brennan, 2001; Shavelson & Webb, 1991).

CONCLUSION

In conclusion, the ITEF has been found in this study not to be unidimensional, but highly reliable (stable across three instances of lesson delivery and dependable). It is only from the perspective of 'use' that the ITEF can be said to measure a single latent construct, which is teaching skill, but it must be known that, this is not from the point of view of statistical tests.

Some limitations in this study must however, be acknowledged. In both the dimensionality and reliability investigation, the rater facet was not included, but treated as an unmeasured facet. This undoubtedly led to the swelling up of the proportion of total variance for the residual in a number of cases of analyses. For example, it was 60.8% for Social Science in the 2018/2019 academic year alone. Future investigations on the ITEF should adopt study designs that will include the rater facet to offer deeper insight into the ITEF's dimensionality and reliability. Again, I cannot rule out completely that no learning occurred between successive evaluations of teaching practice, even though, I tried to curb this by collecting ITEF scores of study subjects that were obtained from successive ratings at very close intervals. Lastly, it must also be stressed that considering what goes into achieving a given level of reliability for a measurement procedure using a particular observation instrument, it is disingenuous to describe the reliability of specific measurement instruments without including the procedures which make the instrument work at the given level of efficiency. Instrument reliability must therefore, be described in cognisance with the instrument, rater training, and specific scoring designs that constitute the measurement procedure.

References

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Prentice Hall/Pearson Education.

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.

Brion, C., & Cordeiro, P. A. (2019). Lessons learned from observing teaching practices: The case of Ghana. J*ournal of Education and Practice, 10*(12).

Brown, G. P., Delgadillo, J., & Golino, H. (2023). Distinguishing the Dimensions of the Original Dysfunctional Attitude Scale in an Archival Clinical Sample. *Cognitive Therapy and Research, 47*(1), 69-83.

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. ERIC.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. Cronbach, L. J. (1972). The dependability of behavioral measurements. *Theory of generalizability for scores and profiles*, 1-33.

Darling-Hammond, L. (2012). *Powerful teacher education: Lessons from exemplary programs*. John Wiley & Sons.

Durvasula, S., Netemeyer, R. G., Andrews, J. C., & Lysonski, S. (2006). Examining the cross-national applicability of multi-item, multi-dimensional measures using generalizability theory. *Journal of International Business Studies, 37*, 469-483.

Finch, W. H., & French, B. F. (2018). *Educational and psychological measurement*. Routledge.

Fortin, A., & Legault, M. (2014). Development of generic competencies: Impact of a mixed teaching approach on students' perceptions. In *Personal Transferable Skills in Accounting Education* (pp. 91-120). Routledge.

Galvão, P. P. d. O., Valente, J. Y., Millon, J. N., Melo, M. H. S., Caetano, S. C., Cogo-Moreira, H., Mari, J. J., & Sanchez, Z. M. (2021). Validation of a tool to evaluate drug prevention programs among students. *Frontiers in psychology, 12*, 678091.

Good, T. L., & Lavigne, A. L. (2015). Rating teachers cheaper, faster, and better: Not so fast. *Journal of Teacher Education, 66*(3), 288-293.

He, L., & Min, S. (2024). Item response theory. In *Development and Validation of a Computerized Adaptive EFL Test* (pp. 15-24). Springer.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational researcher, 41*(2), 56-64.

Hoyt, W. T., & Melby, J. N. (1999). Dependability of measurement in counseling psychology: An introduction to generalizability theory. *The Counseling Psychologist, 27*(3), 325-352.

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.

Irribarra, D. T., & Arneson, A. E. (2023). The challenge of defining and interpreting dimensionality in educational and psychological assessments. *Measurement, 221*, 113430.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151-160.

Lamm, K. W., Lamm, A. J., & Edgar, D. (2020). Scale development and validation: Methodology and recommendations. *Journal of International Agricultural and Extension Education, 27*(2), 24-35.

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.

Li, M., Shavelson, R. J., Yin, Y., & Wiley, E. (2014). Generalizability theory. *The encyclopedia of clinical psychology*, 1-19.

Linacre, J. M. (2023). Advancing the metrological agenda in the social sciences. *Person-Centered Outcome Metrology*, 165.

Luecht, R. M., & Hambleton, R. K. (2021). ITEM RESPONSE THEORY. *The History of Educational Measurement: Key Advancements in Theory, Policy, and Practice*.

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

McDonald, R. P. (2013). *Test theory: A unified treatment*. psychology press.

Micots, C. (2015). Age of elegance: an Italianate sobrado on the Gold Coast. *African Studies Quarterly, 16*(1), 1.

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.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational researcher, 38*(2), 109-119.

Pradhan, R. K., & Hati, L. (2022). The measurement of employee well-being: development and validation of a scale. *Global Business Review, 23*(2), 385-407.

Reckase, M. D. (1990). Unidimensional Data from Multidimensional Tests and Multidimensional Data from Unidimensional Tests.

Reichenheim, M. E., Hökerberg, Y. H. M., & Moraes, C. L. (2014). Assessing construct structural validity of epidemiological measurement tools: a seven-step roadmap. *Cadernos de Saúde Pública, 30*, 927-939.

Rwegoshora, H. M. M. (2016). *A guide to social science research*. Mkuki na Nyota publishers.

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

Smit, N., van de Grift, W., de Bot, K., & Jansen, E. (2017). A classroom observation tool for scaffolding reading comprehension. *System, 65*, 117-129.

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

Surhone, L. M., Timpledon, M. T., & Marseken, S. F. (2010). *Spearman-Brown Prediction Formula*. VDM Publishing. https://books.google.com.gh/books?id=XcRxYgEACAAJ

Timmerman, M. E., Lorenzo-Seva, U., & Ceulemans, E. (2018). The number of factors problem. *The Wiley handbook of psychometric testing: a multidisciplinary reference on survey, scale and test development*, 305-324.

Trumpp, C., Endrikat, J., Zopf, C., & Guenther, E. (2015). Definition, conceptualization, and measurement of corporate environmental performance: A critical examination of a multidimensional construct. *Journal of Business Ethics, 126*, 185-204.

Van der Lans, R. M., Van de Grift, W. J. C. M., & van Veen, K. (2018). Developing an instrument for teacher feedback: using the rasch model to explore teachers' development of effective teaching strategies and behaviors. *The journal of experimental education, 86*(2), 247-264.

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

Wragg, T. (2011). *An introduction to classroom observation (Classic edition)*. Routledge.

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*, 21-29.