frontiers | Frontiers in Education

TYPE Systematic Review
PUBLISHED 08 February 2024
DOI 10.3389/feduc.2024.1329734

OPEN ACCESS

EDITED BY
Hamidreza Namazi,
Monash University Malaysia, Malaysia

REVIEWED BY
Karla Andrea Lobos,
University of Concepcion, Chile
Gregory Siy Ching,
National Chengchi University, Taiwan

*CORRESPONDENCE
John Elvis Hagan Jr.
✉ elvis.hagan@ucc.edu.gh

RECEIVED 29 October 2023
ACCEPTED 26 January 2024
PUBLISHED 08 February 2024

CITATION
Quansah F, Cobbinah A, Asamoah-Gyimah K and Hagan JE Jr. (2024) Validity of student evaluation of teaching in higher education: a systematic review. *Front. Educ.* 9:1329734. doi: 10.3389/feduc.2024.1329734

Validity of student evaluation of teaching in higher education: a systematic review

Frank Quansah[1], Andrews Cobbinah[2], Kenneth Asamoah-Gyimah[2] and John Elvis Hagan Jr.[3,4]*

[1]Department of Educational Foundations, University of Education, Winneba, Ghana, [2]Department of Education and Psychology, University of Cape Coast, Cape Coast, Ghana, [3]Department of Health, Physical Education and Recreation, University of Cape Coast, PMB, Cape Coast, Ghana, [4]Neurocognition and Action-Biomechanics-Research Group, Faculty of Psychology and Sports Science, Bielefeld University Postfach, Bielefeld, Germany

Introduction: Data obtained from students regarding the quality of teaching are used by higher education administrators to inform decisions concerning tenure, promotion, course development and instructional modifications, among others. This article provides a review regarding studies conducted to examine the validity of student evaluation of teaching, specifically focusing on the following objectives: (1) identify the context where studies have been conducted on student evaluation of teaching; (2) find out the methodologies usually employed for assessing the validity of student evaluation of teaching; and (3) establish the sources of measurement error in student evaluation of teaching.

Methods: The systematic review was conducted based on the PRISMA checklist. The databases searched include Scopus, Web of Science (WoS), Google Scholar, PubMed, MEDLINE, ERIC, JSTOR, PsycLIT, EconLit, APA PsycINFO and EBSCO using some specific keywords. After applying the four eligibility criteria, 15 papers were left to be analyzed.

Results: It was discovered that the generalizability theory approach was mostly used to understand the validity of student evaluation data. The review revealed that students were found at the centre of inconsistencies in the evaluation process.

Discussion: The general impression from the review is that the credibility and validity of teaching evaluation outcomes is questionable, considering the several sources of errors revealed. The study recommended closely studying these sources of errors (e.g., rating behaviours of students).

KEYWORDS

student evaluation, higher education, validity, teacher, student, courses

Introduction

Due to the intense competition existing among higher education institutions, universities today are undergoing a paradigm shift in students' status (Raza et al., 2010) such that students are now freely making decisions regarding the type of institution to attend, the programme to choose, and even the type of major courses to read, just like customers selecting their preferred commodities in a supermarket with several varieties of products available (Raza and Khawaja, 2013). Given the significant role students play in the sustenance and running of higher education institutions, they are allowed to evaluate the quality of instruction and courses in their respective institutions. This exercise has become a common phenomenon in almost every university around the globe (Rantanen, 2013).

Student evaluation of teaching is a fairly new concept which was introduced and utilised interchangeably with numerous expressions such as students' appraisal of teaching effectiveness (Marsh, 2007), student appraisal of instructor performance (Chuah and Hill, 2004), students' evaluation of educational quality (SEEQ) (Lidice and Saglam, 2013), student course satisfaction (Betoret, 2007), students' evaluation of instruction (Clayson et al., 2006), or student course evaluation (Chen, 2016; Duggan and Carlson-Bancroft, 2016). Notwithstanding the disparities in the concepts, they have a common underlying objective. Based on the definition and classification of fundamental higher education concepts outlined by the United Nations Educational, Scientific and Cultural Organization (UNESCO), students' appraisal of teaching has been explained as the process by which students assess the general teaching activities, attitude of instructors, and course contents relative to their learning experiences. It is important to emphasize that the critical features evaluated by students include the instructors' ability to explain issues to students, guide their learning, and drive class discussions. The instructors' competencies in unpacking course contents and making courses relevant to students are also key in the assessment process (UNESCO, as cited in Vlăsceanu et al., 2004).

Arguably, student evaluation of courses and teaching influences tenure and promotion (Kogan et al., 2010; Galbraith et al., 2012), students' university application (Alter and Reback, 2014), and student's choice of courses (Wilhelm, 2004). In some advanced countries like the United States, students' evaluation data are used for official and unofficial ranking of institutions, and auditing purposes (Johnson, 2000). These uses of the data have sparked extensive scientific literature in areas such as psychology, education, economics and sociology (Goos and Salomons, 2017). With this understanding, the central issue baffling researchers is the extent to which students' evaluation data can be understood as a pointer for examining the quality of teaching in higher education (Taut and Rakoczy, 2016).

Validity theory and teaching evaluation

Validity in teaching evaluation is the soundness of the interpretation and uses of teaching evaluation results; it is a matter of degree and all about bringing together pieces of evidence to support inferences made about the responses provided by the students (Brookhart and Nitko, 2019). Thus, data from students' appraisal of teaching quality is highly valid when several pieces of evidence can be provided with regard to how the data were taken, the meaningfulness of the data and how the data were used (Brookhart and Nitko, 2019). For validity to be understood in this context, these possibilities should be looked at: (1) raters may not be accurate in their ratings such that the scores given may not reflect the skills/abilities of the instructors or overall teaching quality; (2) the evaluation items may not be clear enough for raters to understand; and (3) students may rate other characteristics of the course and instructor other than actual psychological construct being rated. These issues, among others, are likely to influence the fairness of the evaluation exercise and the results thereof may not be a true reflection of the construct being measured.

In practice, when students are requested to evaluate learning experiences, courses and teaching quality, there is a higher likelihood of disagreement among themselves due to many systematic and unsystematic factors (Eckes, 2015). The systematic factors include differences in rater behaviours, the difficulty of items, and the central tendency effect. On the other hand, the unsystematic factors comprise variations in physical scoring or testing conditions, attention fluctuations of raters as a result of fatigue, errors in transcription, and several others (Brennan, 2011). Whereas these systematic factors produce systematic errors, the unsystematic factors produce random errors (Eckes, 2015). In fact, random (unsystematic) errors can easily be corrected through carefully planned assessment procedures. The case happens to be different when systematic errors are present. Systematic errors, unlike random errors, can be easily identified in a data set. It is worth noting that random errors cancel out in large sample sizes and thus, do not usually have an effect on the validity of large data sets (Eckes, 2015). Systematic errors, on the other side, create a pattern in the data set and distort the meaning derived from the data.

In most evaluation situations, giving a score to depict the degree to which a particular trait is possessed by the object of measurement (i.e., lecturers) is largely based on the subjective judgement of the rater (i.e., student). Rather than operating on collective grounds, student raters regularly seem to considerably differ regarding deeply fixed, more or less individualised rating predispositions, thereby threatening the validity of the evaluation outcomes (Eckes, 2015). Raters also have the probability of giving similar scores for a particular lecturer on a theoretically different criterion (Barrett, 2005; Iramaneerat and Yudkowsky, 2007). Due to this, several raters (as in the case of students' appraisal of instruction), in most cases, are utilised to cancel out random rater errors in the data obtained (Houston and Myford, 2009). The intent of using numerous raters is to have an estimate of every lecturer's teaching quality, which is independent of some specific attribute of the raters. This is done by measuring the attributes of the raters and utilising such data to eliminate the faults of individual raters from the final score of the rater (Linacre, 1994).

Furthermore, raters usually carry out the teaching appraisal by using items with rating scales where points on the scale are required to signify, a consecutively higher degree of performance on the construct (i.e., teaching). It is instructive to mention that most universities use instruments that rely on the analytic rating approach, where raters lookout for specific features of the construct of interest and a score is assigned accordingly based on the extent of the construct's existence. In the case of using the holistic rating method, raters evaluate the whole performance and a single score is assigned to the performance (Engelhard, 2011). With each scoring type, raters are required to differentiate between scale points and assign a rating that appropriately matches the performance; if this is not properly done, the validity of the score assigned will be threatened (Eckes, 2015). This can be a source of measurement error, thereby, serving as a threat to validity.

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

evaluation context, which is a part of the rater-mediated assessment, the major proximal facets include instructors, raters, occasion, and rating items (Brennan, 2011; Eckes, 2015).

Psychometric models for testing teaching evaluation outcomes

The measurement literature identifies three key measurement theories for testing the psychometric analysis of rater-mediator assessments, which includes teaching evaluation by students. These models include the Classical Measurement Theory (CMT), Generalizability Theory (GT), and Item Response Theory (IRT). The CMT stipulates that an actual/observed score rating (X) is a linear model containing a true score (T) and error score (E) [$X = T + E$] (Lord and Novick, 1968). Unlike the observed score, the true score and the error score components are unobserved, which requires some assumptions. In connection with teaching evaluation, any rating score provided by a rater/student is an observed score with two parts; the true score which signifies the precise rating and the error score which denotes the imprecise component of the score. The CMT framework adopts statistical procedures such as regression, correlation and factor analysis with specific methodological designs (e.g., test-retest) to assess whether observed scores are devoid of measurement errors or not (Feldt and Brennan, 1989). It is therefore common to see researchers who obtain teaching evaluation ratings from students on two similar occasions and using correlation analysis to examine the consistency of ratings. In this situation, a high correlation coefficient depicts a high degree of rater consistency and consequently, little error in the rating.

The GT is a statistical theory and conceptual framework for assessing the dependability of a set of observed scores for a specified degree in diverse universes (Cronbach et al., 1963; Shavelson and Webb, 1991). The GT, which has its foundation in the CMT, discards the idea of a single undistinguishable error of measurement and somewhat postulates that the error of measurement occurs from multiple sources (Brennan, 2011) (i.e., $X = T + E_1 + E_2 \ldots + E_x$). These multiple sources of random measurement errors (also known as facets) which inflate construct-irrelevant variances may include raters (i.e., students in this context), rating items and occasions that can be estimated simultaneously within a single analysis (this is a weakness of the CMT) (Eckes, 2015). Unlike the CMT, the GT has the capacity to evaluate the interaction between a number of facets and how this interaction(s) contributes to the variability in observed ratings. For example, how students in a particular class systematically rate lecturers' teaching on a particular item can be explored. Specifically, GT estimates the amount of each error source distinctly and offers an approach for improving the dependability of behavioural measurements. In a statistical sense, the GT combines CMT statistical procedures with the analysis of variance (ANOVA) (Brennan, 2011).

The IRT is a family of psychometric models that provide information about the characteristics of items on an instrument, the persons responding to these items and the underlying trait being measured (Yang and Kao, 2014). Within the context of teaching evaluation, the Many-Facet Rasch Modelling (MFRM), an extended form of the Rasch model under the IRT family, is more appropriate approach to testing the dependability of students' responses (Rasch, 1980; Linacre, 1989). The MFRM allows the evaluation of the characteristics of individual raters, the items and how these raters influence the process of rating (McNamara and Knoch, 2012). MFRM includes the evaluation of the influence of other sources of non-random errors like unreliable raters, inconsistency in ratings across occasions, and inconsistencies in the comparative difficulty of the items (Linacre, 1994). For instance, the MFRM can provide information on whether a single rater tends to systematically score some category of individuals differently than the others, or whether some particular group of individuals performed systematically different on a specific item than they did on others (Linacre, 2003).

Compared to the CMT, the GT and MFRM have both been found in the literature to be more appropriate in terms of assessing the dependability of observed data by estimating the true score and error score components of observed data, especially with rater-mediated assessment or evaluation (Linacre, 2003). The utilisation of these two modelling approaches permits the identification of the various sources of measurement error associated with the observed data. The GT and MFRM procedures, again, compute for reliability indices, which provide an idea of the dependability of the observed data (Brennan, 2011). Moreover, while CMT and GT see the data principally from a group-level viewpoint, separating the sources of measurement error and calculating their extent, an MFRM analysis largely concentrates on individual-level information and, therefore, encourages functional examination into the functioning, or behaviour, of every individual component of the facets being considered (Linacre, 2001).

Rational for the review

The validity of student evaluation of courses and teaching is a contentious issue (Hornstein, 2017). While there are conceptual, theoretical and empirical supports for the validity of students' appraisal of teaching, such data have been critiqued for several reasons (Spooren et al., 2013). In particular, students have been found to evaluate instructional quality based on the characteristics of the course (e.g., difficult/easy nature of courses), student characteristics (e.g., students being friends with the instructor or dislike for the course) and teacher (e.g., the strictness of the instructor) which are unrelated to the quality of teaching and course contents (Berk, 2005; Isely and Singh, 2005; Ko et al., 2013). Such contamination in the evaluation process has several implications for the quality of teaching and learning, instructor growth, tenure and promotion (Ewing, 2012; Galbraith et al., 2012).

The uncertainties surrounding the validity of student evaluation of teaching are partly due to the diverse methodologies employed by the researchers in the field. The measurement literature offers three major approaches (i.e., the CMT, GT and MFRM) that have been used for investigating the dependability of student evaluation of teaching. The question of which methodology provides the most comprehensive and accurate results in performance-mediated assessment has been extensively discussed in the measurement literature. In addition to what has been discussed earlier on, two conclusions have been made: (1) GT is preferred to CMT because the weaknesses of CMT have been curtailed by GT (see Cronbach et al., 1972; Shavelson and Webb, 1991; Brennan, 2001b), (2) MFRM as compared to GT offers more knowledge about the data, but combining both approaches offers an excellent picture of the rating process and supports the current idea of validity evidence (Linacre, 2003; Kim and Wilson, 2009; Brennan, 2011; Lee and Cha, 2016).

This review aims to offer a systematic imprint of literature on the dependability of student evaluation of teaching in higher education. Similar reviews have been conducted on the issue of the usefulness and validity of student appraisal of teaching (Costin et al., 1971; Wachtel, 1998; Onwuegbuzie et al., 2006; Spooren et al., 2013); however, these studies were narrative and scooping in nature and largely focused on distal factors such distal (e.g., gender, age, attitude, ethnicity, motivation, etc), rather than proximal factors (e.g., item difficulty, rater severity, rating scale functioning, etc)[1] which is the focus of this review. The recent review conducted in 2013 by Spooren et al., for example, discussed validity issues in questionnaire design, dimensionality, online evaluation, teacher characteristics, and particularly, content and construct validity. In fact, none of the earlier reviews studied the methodologies, sources of variability in terms of proximal factors, and the context in which studies have been carried out. These dimensions are necessary ingredients in terms of understanding the teaching evaluation landscape in higher education, appropriately driving professional practice, policy formulation and implementation. Unlike previous reviews, this study was aimed at conducting a systematic review to achieve the following objectives: (1) To identify the context where studies have been conducted on student evaluation of teaching; (2) To find out the methodologies usually employed for assessing the validity of student evaluation of teaching; and (3) To understand the sources of measurement error in student evaluation of teaching.

This paper is significant for some reasons. First, the outcome of this review would enlighten administrators of higher education institutions on the sources of measurement errors and the extent of dependability of student evaluation data. This information will help these administrators on how the teaching evaluation can be carried out in an error-free setting and at the same understating the extent to which the outcome of the evaluation can be utilised. Secondly, this review would inform the direction of further studies in terms of (a) the methods future researchers should adopt in their investigation; (b) which study settings need more attention in terms of research; and (c) which specific factors should be well studied to understand the variability in student evaluation of teaching.

Methods

Research protocol

This systematic review was conducted based on the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 extended checklist (Page et al., 2021). All protocols followed in this research were guided by PRISMA.

Search plan and information sources

The databases searched included Scopus, Web of Science (WoS), Google Scholar, PubMed, MEDLINE, ERIC, JSTOR, PsycLIT, EconLit, APA PsycINFO and EBSCO. The search for the literature was done by three independent researchers and carried out within the period, 15th to 26th December, 2022. The search was conducted by combining these three keywords (i.e., "validity," "reliability," and "variability") with each of these phrases (i.e., "student evaluation of teaching," "teaching quality appraisal," "student appraisal," and "teaching evaluation"). The Boolean operator "and" was used for combining the keywords and phrases. The language filter was applied to restrict the search to all manuscripts written in the English Language. After the initial search, 293 papers were retrieved. There were no year restrictions used for deciding which paper was eligible or not. Duplicates were detected through the Zotero tool. Some other duplications were also deleted manually. In all, 41 duplicates were deleted.

Screening procedure

Two hundred and fifty-two (252) papers were independently screened by three researchers with the following educational background and expertise: psychology, programme evaluation, measurement in education and psychology, psychometrics, and research methodology. First, the papers were screened by critically considering the titles and abstracts focusing on quantitative studies. After this phase, 113 papers were exempted and the remaining 139 papers were further screened for eligibility. The following criteria were set for the exclusion of papers for the analysis:

1. Articles which investigated the validity of the student evaluation of teaching using a statistical (quantitative) approach under CMT, GT and IRT, and not based on opinions of students, regarding the quality of such data were included. We focused on quantitative studies because it is the only means by which rating inconsistent rating beahaviours can be directly observed. Qualitative studies can only take the opinions of students, lecturers, and other stakeholders concerning the validity of the outcome of teaching evaluation. Whereas these opinions may be useful, such views may not be objective but based on the inter-subjectivity experiences of the respondents;
2. Studies which were conducted on the dependability of student appraisal of teaching in higher education. The study focused on higher education because there has not been consensus on the measurement of the quality of teaching at the pre-tertiary level of education (Chetty et al., 2014);
3. Studies focusing on how the distal factors contribute to the variations in student appraisal of teaching were also excluded. For example, articles that conducted factor analysis or questionnaire validation were excluded for two reasons: (1) several reviews have been conducted on the development and validation of appraisal questionnaires; (2) almost every higher education institution validates the questionnaire for the evaluation exercise. This was supported in the studies which

1 Distal facets are those variables that may have a mediated or indirect influence on the ratings provided by students (Eckes, 2015). Unlike distal facets, proximal facets have an immediate and direct effect on the rating scores awarded to the lecturers. Thus, the proximal facets, especially the person facet, plays a significant role in understanding the validity dynamics of the evaluation data (Eckes, 2015).

were analyzed for this paper; all the studies used a well-validated instrument for data collection.

During the last screening phase, which is the application of the eligibility criteria, a more detailed reading was done by all four researchers. In the end, 124 out of 139 papers failed to meet the eligibility criteria, leading to a final sample of 15 papers that were analyzed and synthesized for the study (see Figure 1).

Data analysis plan

Data extraction is based on the 15 papers by coding the information into the following themes: (1) author(s) and publication year, (2) country where the study was conducted, (3) statistical approach adopted by the author(s), (4) the proximal factors understudied, and (5) the key idea from the research (see Table 1). To check the appropriateness of the coding, all four investigators were involved in the process; this was based on the recommendations of González-Valero et al. (2019). Discrepancies were resolved among the investigators but most importantly, the themes identified by the investigators were consistent. To ensure sufficient reliability, inter-rater agreement was calculated based on the Fleiss' Kappa (Fk) statistical index. A coefficient of 0.77 was achieved for the information extraction and selection, which indicated adequate agreement (Fleiss, 1971).

Results

The context of the studies

The review showed that research works on the sources of variability and the validity of student evaluation of teaching and courses in schools have been extensively conducted across North America and Asia. All the studies on the North American continent

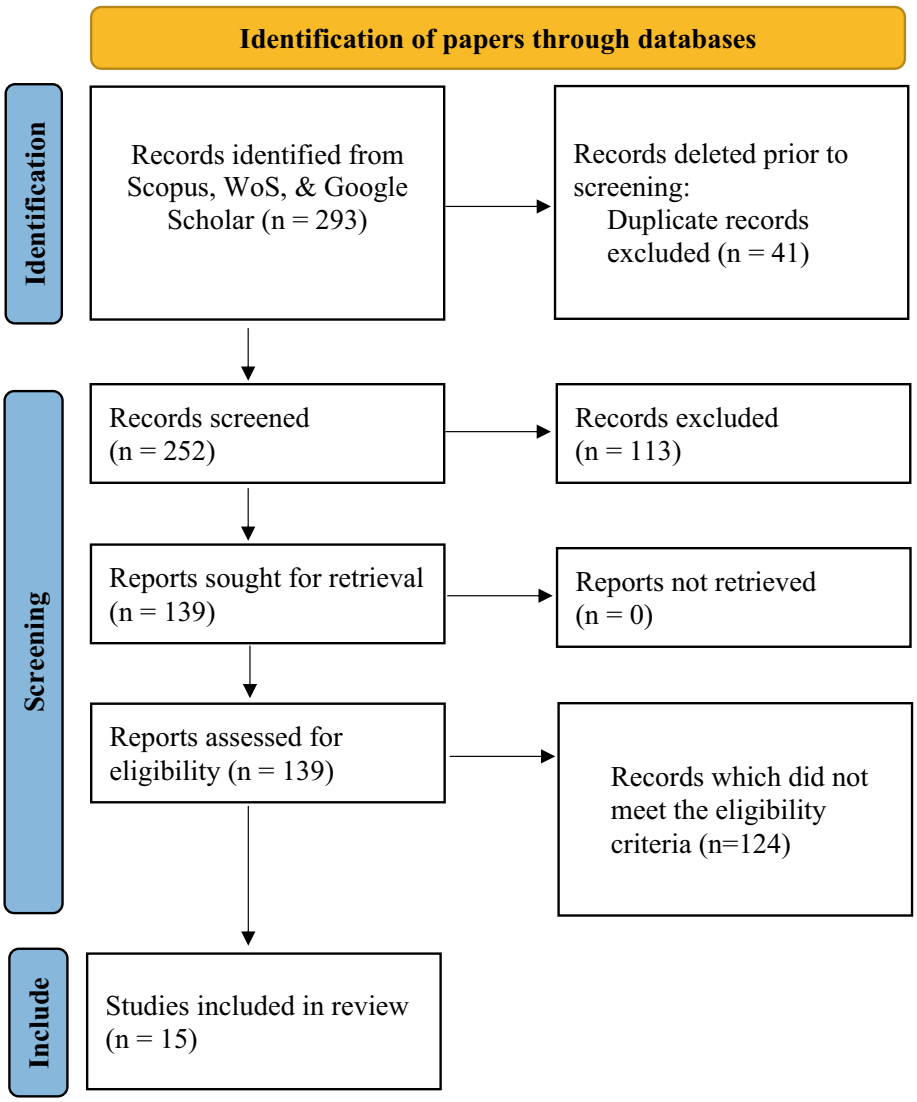

FIGURE 1
PRISMA flow chart.

TABLE 1 Summary of studies on student evaluation of teaching in higher education.

Author(s) and year	Country	Approach	Source of variability	Key idea
Gillmore et al. (1978)	Washington, USA	GT	Student, Instructor*, Item, & Course	The student (nested within teachers/courses) contributes the largest variability in student evaluation of teaching. The course facet did not play any significant role in the variations in student ratings. Item facet had the least contribution
Ibrahim (2011)	Sultanate of Oman	GT	Student, Item, & Instructor*	Students (nested within instructors), and instructor-by-items-by-student interaction had a large variance contribution. Item facet had the least contribution
Kane et al. (1976)	Illinois, USA	GT	Student, Item, & Class	Students (nested within instructors), and instructor-by-items-by-student interaction had a large variance contribution. Item facet had the least contribution
Marsh and Overall (1980)	California, USA	CMT	Student	Lack of consistency was found in the ratings of the students
Mazor et al. (1999)	Massachusetts, USA	GT	Student, Item, & Instructor*	The students (nested in teacher) had the largest variance contribution Item facet had the least contribution
Samian and Noor (2012)	Malaysia	CMT	Student	A high rate of correlation was found between the comments and the ratings. The study established that students' ratings of the lecturer's teaching quality were highly valid and reliable
VanLeeuwen et al. (1999)	New Mexico, USA	GT	Item, Class, & Student	The residual (interaction of all the facets) contributed the largest variances in the ratings of students High level of validity of students' rating
Feistauer and Richter (2016)	Germany	GT	Student, Course, & Instructor*	The results suggested a flaw and cast some uncertainty on the validity of student rating of teaching quality The interaction of teachers and students was the largest source of variance
Rindermann and Schofield (2001)	Germany	CMT	Student	There was consistency in the ratings of students across courses handled by the same instructor
Börkan (2017)	Turkey	MFRM	Student, Item, & Instructor	It was revealed that the students largely differed in the level of severity while rating instructors. The findings of the study questioned the validity of students' ratings as students failed to evaluate teachers as was expected of them
Quansah (2020)	Ghana	GT	Student, Item, & Occasion	The residual (interaction of all facets) and students (nested in class) contributed the largest variance to students' ratings. The results found that the overall dependability of students' rating of lecturers was low. Item and occasion had the least contribution
Li et al. (2018)	China	GT	Class & Student	Ratings of students appeared to be inconsistent across time and programme major
Spooren et al. (2014)	Belgium	GT	Student*, Item, & Occasion	Student-by-item interaction had the largest contribution to the variability in student ratings. Item and item-by-occasion had the least contribution
Üstünlüoğlu and Can (2012)	Turkey	CMT	Student	A positive relationship between students' evaluation and coordinators' ratings of the same instructor over the two years was revealed
Quansah (2022)	Ghana	MFRM	Student, Item, & Instructor	Inconsistent rating behaviours, poor item functioning and scale structure, halo effect, and non-functional rating scale were reported in the teaching evaluation exercise

*These facets was not considered as a source of measurement error in the study.

were conducted in the USA, specifically in New Mexico (VanLeeuwen et al., 1999), Washington (Gillmore et al., 1978), Illinois (Kane et al., 1976), Massachusetts (Mazor et al., 1999), and California (Marsh and Overall, 1980). All of these studies in the USA were carried out before the year 2000, which perhaps, indicates a declining interest in the issue of the sources of variability in students' appraisal of teaching. The trend of studies in Asia was different from those studies in the USA (see Table 2). Studies that investigated the sources of variability in student ratings of instructors in Asia started not long ago when Ibrahim (2011) studied that issue in the Sultanate of Oman using the

TABLE 2 The context of previous studies.

Continents	No. of studies	Countries (frequency)	Approaches (frequency)	Years
North America	5	USA (5)	GT (4), CMT (1)	1976, 1978, 1990,1999, 1999
Asia	5	Turkey (2), Malaysia (1), China (1), Oman (1)	GT (2), CMT (2), MFRM (1)	2011, 2012*, 2017, 2018
Europe	3	Belgium (1), Germany (2)	GT (2), CMT (1)	2001, 2014, 2016
Africa	2	Ghana (2)	GT (1), MFRM (1)	2020, 2022

*There were two studies for that year.

GT approach. Since then, studies have been conducted by Samian and Noor (2012) in Malaysia, Börkan (2017), and Üstünlüoğlu and Can (2012) all in Turkey, and Li et al. (2018) in China. A few of the studies were conducted in Europe ($n = 3$, 20%) (Rindermann and Schofield, 2001; Spooren et al., 2014; Feistauer and Richter, 2016) and, particularly, in Africa ($n = 2$, 13.3%) (Quansah, 2020, 2022). Although few studies were carried out in Europe and Africa, most of the studies were quite current compared to those carried out in America.

Methodology utilized

The review revealed that all three measurement theories (i.e., CMT, GT, and MFRM) have been utilized to investigate the dependability of student ratings of lecturers' teaching and courses. However, the GT approach was found to be more popular in terms of its usage. Out of the 15 empirical studies which met the criteria for this review, 9 (60%) of them used the GT approach for their investigation (see Table 1). Comparatively, the use of the GT approach appears to be gaining ground and dominating the quality assurance in higher education literature in recent times (see Figure 2). Additional four studies (26.7%) used CMT and only two (13.3%) research works adopted the MFRM approach (see Figure 2) (Börkan, 2017; Quansah, 2022).

Main sources of measurement errors

The review revealed five main sources of variability in student rating exercises. This includes student, evaluation item/scale, occasion, teacher, class, and course. For all the nine studies which adopted GT, the teacher, class or course was used as the object of measurement. In addition, the teacher factor was considered as the measurement object for two studies that adopted the MFRM. The findings on the specific sources of variability have been provided subsequently.

Student

It was found that all the studies reviewed included the student (i.e., rater) as a source of variability in teaching evaluation rating (see Table 1). However, one of the studies (Spooren et al., 2014) failed to recognize the student facet as a source of error in the ratings because the authors believed that student variability is desired since it represents disparities in individual students in their quality ratings of a course. Despite this argument by the authors (Spooren et al., 2014), it was revealed that the students were inconsistent with how they responded to the items. Those studies which considered students as a source of measurement error had three common findings; (1) student was the only main effect variable that recorded the largest contribution to the variability in the ratings; (2) the analysis showed a low level of validity of the responses provided by the students in rating instructors/courses; and (3) increasing the number of students who participated in the evaluation for each instructor was more useful to improving the validity of student ratings.

In a more detailed analysis using the MFRM, two studies (Börkan, 2017; Quansah, 2022) revealed that although some students provided accurate responses during the teaching evaluation exercise, quite a number of them were inconsistent with rating their instructors. Other wide-range of rating behaviours reported in these two studies include: (1) some students failed to discriminate between the quality of teaching or course quality or the performance level of the instructor(s), indicating that these students provided similar ratings for completely different situations (i.e., lecturers with different teaching ability); (2) the majority of the students were influenced by (less salient) factors which are unrelated to the targeted teaching behaviours they are required to assess. Take for example, a lecturer who is punctual to class yet has weak subject matter and pedagogical knowledge; most students provided excellent ratings for instructors being influenced by the lecturer's punctual behaviours even at points when punctuality is not assessed; (3) the majority of the students were lenient in their ratings, providing high ratings for undeserving instructors or situations.

Evaluation items/scale

Nine (60%) out of the 15 studies considered the item as a major source of variance in student ratings. Out of the nine studies, seven of them found that the item had the least variance contribution indicating that the evaluation item contributed very few measurement errors to the quantification of the ratings (see Kane et al., 1976; Gillmore et al., 1978; Mazor et al., 1999; VanLeeuwen et al., 1999; Ibrahim, 2011; Spooren et al., 2014; Quansah, 2020). The remaining two studies (Börkan, 2017; Quansah, 2022) investigated beyond the item-variance properties to the scale functioning using MFRM. In their research, these authors revealed poor scale functioning with most responses clustered around the highest two scale categories for a 5-point scale coupled with a lack of clarity of the response categories. Moreover, some of the items on the evaluation instruments were identified as unclear, redundant and could not measure the targeted trait (Börkan, 2017; Quansah, 2022).

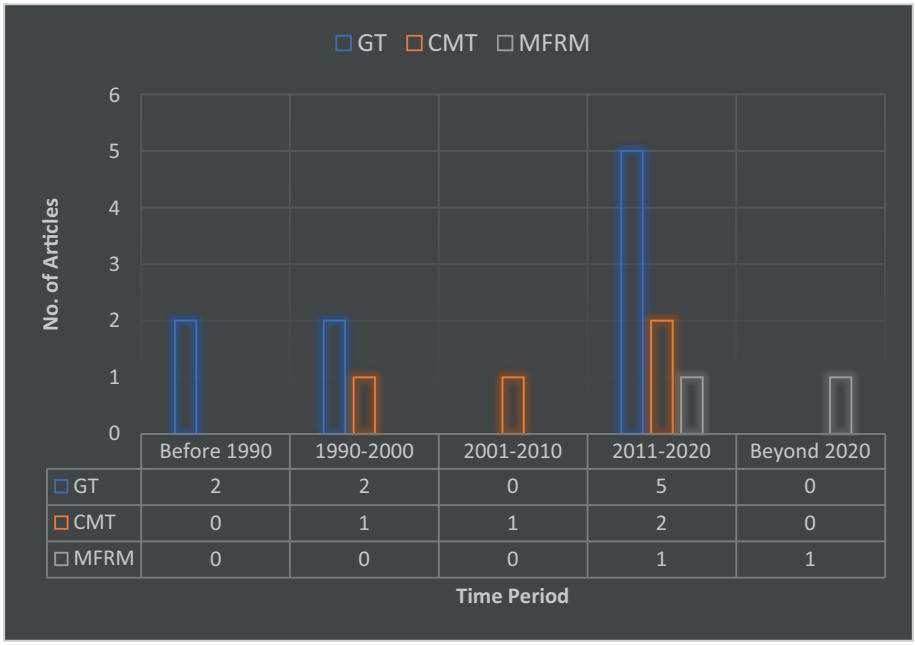

FIGURE 2
Methods used in previous studies over time.

Occasion

Two of the studies (Spooren et al., 2014; Quansah, 2020); these studies adopted the GT methodological approach. Findings from both studies revealed that occasion did not contribute significantly to measurement errors during the ratings. It was further revealed that one-time evaluation data offered much more advantage in terms of precision in student responses than obtaining the evaluation data for a lecturer on multiple occasions from the same group of students. However, one of the studies (i.e., Quansah, 2020) indicated that taking the evaluation data in the middle of the semester yielded a more accurate response from the students than waiting until the semester ends.

Instructor/class/course

Few studies considered instructors, class and course type as sources of measurement error (Gillmore et al., 1978; Feistauer and Richter, 2016; Börkan, 2017; Quansah, 2022). These studies found that class and course type contributed very little in terms of the variances in the ratings of students. For instructors, two of the studies revealed that the instructors did not receive accurate and precise ratings from the students. In most cases, the instructors received higher ratings than expected.

Credibility of students' evaluation data

Generally, the majority of the studies demonstrated that the validity of student evaluation data was low. Whereas the studies which utilized GT and MFRM approaches showed a low level of validity of student ratings of instructors, the majority of the studies which adopted the CMT methodology found a high level of validity of the data. Except for Marsh and Overall (1980) who found a lack of consistency in the ratings by students in a university in Los Angeles. Studies by Rindermann and Schofield (2001), Samian and Noor (2012), and Üstünlüoğlu and Can (2012) revealed a high validity level of student ratings. What is found common with all the studies that adopted the CMT is that they all employed the criterion validity approach, where they attempted to corroborate statistical results from one occasion to another through correlational analysis. These researchers appeared to focus on the stability of traits or triangulation of evaluation.

Discussion

The purpose of this review was to analyse and synthesize existing studies on the subject of the validity of students' evaluation of teaching across the globe. The review attempted to understand the scope and context of the research area regarding the subject, the methodologies adopted by the available evidence and the factors that account for errors in student responses during teaching evaluation. The outcome of the review showed that for over five decades, the available literature on the validity of teaching evaluation data is scanty, with the majority of the recent studies conducted in Asia and North America while a few have been conducted in Europe and Africa. Interestingly, the recent interest in the subject area was found among researchers in Europe; meanwhile, scholars in North America since 1999 have not conducted any research in the area. This finding provides insight into how researchers on different continents are directing their research focus to the area. The geographical distribution of studies and the scanty available research can be tied to the complexities in studying the validity of teaching evaluation in higher education due to the

highly statistical approach required (Ashaari et al., 2011; Rosli et al., 2017).

The review showed that the majority of the studies adopted the GT approach to their investigation while just a few utilised the MFRM. The popularity of GT can be explained for two reasons. First, GT was introduced to address some weaknesses in the oldest measurement theory (i.e., CMT) in terms of its use in performance-mediated assessments (Shavelson and Webb, 1991). This led to the switch from CMT to GT, even though this transition took some time due to the complexities in the use of GT such as developing syntaxes for running the analysis (Brennan, 2001b). Thus, it could be observed that studies adopting the GT have grown steadily from 1976 to 2020. This also explains the decreased levels of use of CMT in recent times. Secondly, several computer programming software or syntaxes are available today with their respective instruction or guidelines which have made the adoption of GT less difficult. Some of the software or syntaxes include GENOVA package (Brennan, 2001a), ETUDGEN (Cardinet et al., 2010), MATLAB, SAS, SPSS (Mushquash and O'Connor, 2006), EduG (Cardinet et al., 2010), G-String, LISREL (Teker et al., 2015), and R programming (Hucbner and Lucht, 2019) among others. Thus, the availability of these statistical software and syntaxes can make the use of GT more popular. Concerning the MFRM approach, only two studies adopted the MFRM approach. The low adoption of MFRM analysis in the higher education quality assurance literature can also be attributed to the fact that most scholars are not aware of such a procedure. This coupled with the complexity of the use of the approach/method, which requires a high level of expertise, especially when there are so many factors involved. There is also relatively few computer programme software/syntaxes which can perform MFRM analysis. The two known applications are FACET and R programming applications (Lunz et al., 1990). The programming nature of these software and, perhaps, the limited guidelines for their use may have discouraged researchers who have little background in measurement but are scholars in quality assurance.

A more significant aspect of the findings showed that the student (i.e., the rater) contributed the largest amount of errors to the teaching evaluation data. It was found that higher education students are inconsistent in their ratings with most of them failing to provide ratings that discriminate across the varying levels of performance of instructors. Most students were also influenced by (less salient) factors which are unrelated to the targeted construct being measured. What is central to this finding is the fact that higher education students are not usually trained to respond to teaching evaluations (Eckes, 2015). In many institutions, students are offered a brief orientation about what the teaching evaluation is about without proper training (Dzakadzie and Quansah, 2023). It is not surprising that some students showed a lack of understanding of the response options on the evaluation instrument, although these instruments had excellent psychometric properties. Higher education administrators should organize regular training programmes for students on how to rate accurately to reduce errors of measurement (such as halo effect, inconsistent rating, and inability to use the rating scales) during the teaching appraisal exercise. The training should include what (behaviours) they should look out for when appraising.

Further analysis from the review showed that a greater proportion of tertiary students are lenient in their ratings. Although the reasons for their leniency were not explored in the various studies, some factors are obvious considering the framework of teaching evaluation. A key concern is the issue of negative critical culture where students experience fear of retaliation by the lecturer when they provide poor teaching evaluation. In such instances, students may feel reluctant to share honest opinions about teaching activities and services they receive from the institution (Adams and Umbach, 2012). This negative atmosphere can be worsened when anonymity and confidentiality of student responses cannot be assured. This situation might have contributed to the findings that the instructors received very high evaluation scores. An interesting perspective on this issue is that several pieces of research work have confirmed that grade inflation is positively associated with teaching evaluation outcomes (Eiszler, 2002; Stroebe, 2020; Berezvai et al., 2021; Park and Cho, 2023). The takeaway from these studies is that some professors exchange lenient grading with excellent evaluation results. This reason can explain why one of the studies included in the systematic review showed that the teaching evaluation conducted in the middle of the semester (before any assessment) had higher reliability than those performed at the end of the semester (i.e., when some assessments have been conducted). While higher education institutions are encouraged to uphold anonymity and confidentiality, students should be oriented on why they need to provide honest responses without fear of reprisals. It is also suggested higher education administrators should orient professors/instructors on the benefits associated with accurate evaluation data. Additionally, teaching evaluation should be organised by the authorised department preferably before the end of the semester. Other strategies can be explored to decouple students' grades from their evaluation responses.

The general impression across the available studies from different continents is that the credibility and validity of teaching evaluation outcomes is questionable, considering the several sources of errors revealed. The majority of the evidence from the empirical papers reviewed suggests little support for administrators to rely on teaching evaluation results for critical decisions concerning instructors and policy implications. The outcome of the review draws on a close partnership among students, professors/instructors and management of higher education institutions in ensuring that the reliability of such data is improved. By having a clear framework of responsibilities for all these parties and stakeholders, much progress can be made considering the implications of the evaluation results for all parties. Despite the relevance of this review, it only included quantitative studies (and excluded studies which examined the opinions of stakeholders regarding the validity of teaching evaluation results). We recommend that future researchers are encouraged to conduct a systematic review of qualitative studies conducted on the subject.

Conclusions and future research direction

The results from the review lead the researchers to a general conclusion that not much has been done in exploring the sources of variation and validity of student ratings of teachers/courses in institutions of higher education. This situation calls for an urgent need for more empirical research work to be conducted in the area. These limited studies, however, have drawn upon a universal consensus that the validity of students' responses to the appraisal of teaching in higher education is still in doubt. Students are found at the centre of these inconsistencies in the rating process due to several factors that could

not be disentangled from the students. This paper, therefore, serves as a prompt to researchers to conduct more studies in the area.

We recommend that further studies adopt the MFRM methodological framework or possibly blend these procedures to continue the discussion on the validity of evaluation responses by students. It is essential to note that merging these procedures (especially GT and MFRM) supports recent developments in validity theory which recommends multiple sources of validity evidence to be gathered, assessed, and combined into a validity argument to support score interpretation and utilisation (Kane, 2012; Fan and Bond, 2016). Since students played a pivotal role in terms of understanding the variations in student ratings of teachers/courses in higher education, we recommend that further research should be conducted to closely study the behaviours of students during the evaluation exercise. This investigation should be extended to examining the process data by observing the behavioural patterns of the students in the rating process through log files.

Although few studies examined the scale functioning quality of the evaluation form used, this serves as a prompt for future studies to take a close look at the scale functioning of the evaluation instrument. It must be mentioned that most validation procedures for evaluation forms do not include scale category functioning. Future research should include the item as a source of measurement error for further careful examination including the scale category quality. Except for Asia, future research should be conducted in other continents like Africa, Europe, South America, Antarctica, and Australia. This is essential to provide such information to administrators of higher education regarding the utility of evaluation of teaching data from students. This is also needed to help understand the sources of measurement error, particularly proximal factors, and the validity of student ratings of teachers/courses in higher education around the globe.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

FQ: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. AC: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. KA-G: Investigation, Methodology, Validation, Visualization, Writing – review & editing. JH: Funding acquisition, Investigation, Methodology, Validation, Visualization, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The study received funding from the School of Graduate Studies and Graduate Students Association (GRASSAG) of the University of Cape Coast, the Samuel and Emelia Brew-Butler/SGS/GRASSAG-UCC Research Grant. The authors sincerely thank Bielefeld University, Germany for providing financial support through the Open Access Publication Fund for the article process charge.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Note: Those sources marked with an asterisk were the studies included in the review.

Adams, M. J., and Umbach, P. D. (2012). Nonresponse and online student evaluations of teaching: understanding the influence of salience, fatigue, and academic environments. *Res. High. Educ.* 53, 576–591. doi: 10.1007/s11162-011-9240-5

Alter, M., and Reback, R. (2014). True for your school? How changing reputations alter demand for selective U.S. colleges. *Educ. Eval. Policy Anal.* 36, 346–370. doi: 10.3102/0162373713517934

Ashaari, N. S., Judi, H. M., Mohamed, H., and Wook, M. T. (2011). Student's attitude towards statistics course. *Procedia Soc. Behav. Sci.* 18, 287–294. doi: 10.1016/j.sbspro.2011.05.041

Barrett, S. (2005). "Raters and examinations" in *Applied Rasch measurement: a book of exemplars*. eds. S. Alagumalai, D. C. Curtis and N. Hungi (Dordrecht: Springer), 159–177.

Berezvai, Z., Lukáts, G. D., and Molontay, R. (2021). Can professors buy better evaluation with lenient grading? The effect of grade inflation on student evaluation of teaching. *Assess. Eval. High. Educ.* 46, 793–808. doi: 10.1080/02602938.2020.1821866

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *Int. J. Teach. Learn. High. Educ.* 17, 48–62.

Betoret, F. D. (2007). The influence of students' and teachers' thinking styles on student course satisfaction and on their learning process. *Educ. Psychol.* 27, 219–234. doi: 10.1080/01443410601066701

*Börkan, B. (2017). Exploring variability sources in student evaluation of teaching via many-facet Rasch model. *J. Meas. Eval. Educ. Psychol.*, 8, 15–33, doi: 10.21031/epod.298462

Brennan, R. L. (2001a). "Manual for urGENOVA version 2.1" in *Iowa testing programs occasional paper number 49* (Iowa City, IA: Iowa Testing Programs, University of Iowa).

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

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Appl. Meas. Educ.* 24, 1–21. doi: 10.1080/08957347.2011.532417

Brookhart, S. M., and Nitko, A. J. (2019). *Educational assessment of students*. Upper Saddle River, NJ: Pearson.

Cardinet, J., Johnson, S., and Pini, G. (2010). *Applying generalizability theory using EduG*. New York, NY: Routledge/Taylor and Francis Group.

Chen, L. (2016). Do student characteristics affect course evaluation completion?. Annual Conference of the Association for Institutional Research. New Orleans, LA.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates. *Am. Econ. Rev.* 104, 2593–2632. doi: 10.1257/aer.104.9.2593

Chuah, K. L., and Hill, C. (2004). Student evaluation of teacher performance: random pre-destination. *J. Coll. Teach. Learn.* 1, 109–114. doi: 10.19030/tlc.v1i6.1961

Clayson, D. E., Frost, T. F., and Sheffet, M. J. (2006). Grades and the student evaluation of instruction: a test of the reciprocity effect. *Acad. Manage. Learn. Educ.* 5, 52–65. doi: 10.5465/amle.2006.20388384

Costin, F., Greenough, W. T., and Menges, R. J. (1971). Student ratings of college teaching: reliability, validity, and usefulness. *Rev. Educ. Res.* 41, 511–535. doi: 10.3102/00346543041005511

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). *The dependability of behavioural measurements: theory of generalizability for scores and profiles*. New York, NY: Wiley.

Cronbach, L. J., Rajaratnam, N., and Gleser, G. C. (1963). Theory of generalizability: a liberalization of reliability theory. *Br. J. Stat. Psychol.* 16, 137–163. doi: 10.1111/j.2044-8317.1963.tb00206.x

Duggan, M., and Carlson-Bancroft, A. (2016). How Emerson College increased participation rates in course evaluations and NSSE. Annual Conference of the Association for Institutional Research. New Orleans, LA.

Dzakadzie, Y., and Quansah, F. (2023). Modelling unit non-response and validity of online teaching evaluation in higher education using generalizability theory approach. *Front. Psychol.* 14:1202896. doi: 10.3389/fpsyg.2023.1202896

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: analysing and evaluating rater-mediated assessment* (*2*). Frankfurt: Peter Lang GmbH.

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Res. High. Educ.* 43, 483–501. doi: 10.1023/A:1015579817194

Engelhard, G. (2011). Evaluating the bookmark judgments of standard-setting panellists. *Educ. Psychol. Meas.* 71, 909–924. doi: 10.1177/0013164410395934

Ewing, A. M. (2012). Estimating the impact of relative expected grade on student evaluations of teachers. *Econ. Educ. Rev.* 31, 141–154. doi: 10.1016/j.econedurev.2011.10.002

Fan, J., and Bond, T. (2016). Using MFRM and SEM in the validation of analytic rating scales of an English speaking assessment. In Q. Zhang (Ed.). Conference Proceedings for Pacific Rim Objective Measurement Symposium (PROMS) 29–50.

*Feistauer, D., and Richter, T. (2016). How reliable are students' evaluations of teaching quality? A variance components approach. *Assess. Eval. High. Educ.*, 10, 1–17. Doi: 10.1080/02602938.2016.1261083

Feldt, L. S., and Brennan, R. L. (1989). "Reliability" in *Educational measurement*. ed. R. L. Linn. *3rd* ed (New York: American Council on Education and MacMillan), 105–146.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychol. Bull.* 76, 378–382. doi: 10.1037/h0031619

Galbraith, C. S., Merrill, G. B., and Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business-related classes? A neural network and Bayesian analyses. *Res. High. Educ.* 53, 353–374. doi: 10.1007/s11162-011-9229-0

*Gillmore, G. M., Kane, M. T., and Naccarato, R. W. (1978). The generalizability of student ratings of instruction: estimation of the teacher and course components. *J. Educ. Meas.*, 15, 1–13, doi: 10.1111/j.1745-3984.1978.tb00051.x

González-Valero, G., Zurita-Ortega, F., Ubago-Jiménez, J. L., and Puertas-Molero, P. (2019). Use of meditation and cognitive behavioral therapies for the treatment of stress, depression and anxiety in students. A systematic review and meta-analysis. *Int. J. Environ. Res. Public Health* 16, 1–23. doi: 10.3390/ijerph16224394

Goos, M., and Salomons, A. (2017). Measuring teaching quality in higher education: assessing selection bias in course evaluations. *Res. High. Educ.* 58, 341–364. doi: 10.1007/s11162-016-9429-8

Hornstein, H. A. (2017). Student evaluations of teaching are inadequate assessment tool for evaluating faculty performance. *Cogent Educ.* 4, 13–42. doi: 10.1080/2331186X.2017.1304016

Houston, J. E., and Myford, C. M. (2009). Judges' perception of candidates' organization and communication in relation to oral certification examination ratings. *Acad. Med.* 84, 1603–1609. doi: 10.1097/ACM.0b013e3181bb2227

Hucbner, A., and Lucht, M. (2019). Generalizability theory in R. *Pract. Assess. Res. Eval.* 24, 5–12. doi: 10.7275/5065-gc10

*Ibrahim, A. M. (2011). Using generalizability theory to estimate the relative effect of class size and number of items on the dependability of student ratings of instruction. *Psychol. Rep.*, 109, 252–258, doi: 10.2466/03.07.11.PR0.109.4.252-258

Iramaneerat, C., and Yudkowsky, R. (2007). Rater errors in a clinical skills assessment of medical students. *Eval. Health Prof.* 30, 266–283. doi: 10.1177/0163278707304040

Isely, P., and Singh, H. (2005). Do higher grades lead to favourable student evaluations? *J. Econ. Educ.* 36, 29–42. doi: 10.3200/JECE.36.1.29-42

Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teach. High. Educ.* 5, 419–434. doi: 10.1080/713699176

Kane, M. (2012). Validating score interpretations and uses. *Lang. Test.* 29, 3–17. doi: 10.1177/0265532211417210

*Kane, M. T., Gillmore, G. M., and Crooks, T. J. (1976). Student valuation of teaching: the generalizability of class means. *J. Educ. Meas.*, 13, 173–183, Doi: 10.1111/j.1745-3984.1976.tb00009.x

Kim, S. C., and Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model. *J. Appl. Meas.* 10, 408–423.

Ko, J., Sammons, P., and Bakkum, L. (2013). *Effective teaching: a review of research and evidence*. Berkshire: CfBT Education Trust.

Kogan, L. R., Schoenfeld-Tacher, R., and Hellyer, P. W. (2010). Student evaluations of teaching: perceptions of faculty based on gender, position, and rank. *Teach. High. Educ.* 15, 623–636. doi: 10.1080/13562517.2010.491911

Lee, M., and Cha, D. (2016). A comparison of generalizability theory and many facet Rasch measurement in an analysis of mathematics creative problem-solving test. *J. Curric. Eval.* 19, 251–279. doi: 10.29221/jce.2016.19.2.251

*Li, G., Hou, G., Wang, X., Yang, D., Jian, H., and Wang, W. (2018). A multivariate generalizability theory approach to college students' evaluation of teaching. *Front. Psychol.* 9:1065. doi: 10.3389/fpsyg.2018.01065

Lidice, A., and Saglam, G. (2013). Using students' evaluations to measure educational quality. *Procedia Soc. Behav. Sci.* 70, 1009–1015. doi: 10.1016/j.sbspro.2013.01.152

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1994). *Many-facet Rasch measurement* (*2*). Chicago: MESA Press.

Linacre, J. M. (2001). Generalizability Theory and Rasch Measurement. *Rasch Measurement Transactions*, 15, 806–807.

Linacre, J. M. (2003). *A user's guide to FACETS (computer program manual)*. Chicago: MESA Press.

Lord, F. M., and Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lunz, M., Wright, B., and Linacre, J. (1990). Measuring the impact of judge severity on examination scores. *Appl. Meas. Educ.* 3, 331–345. doi: 10.1207/s15324818ame0304_3

Marsh, H. W. (2007). "Students' evaluations of university teaching: a multidimensional perspective" in *The scholarship of teaching and learning in higher education: an evidence-based perspective* (Dordrecht: Springer), 319–384.

*Marsh, H. W., and Overall, J. U. (1980). Validity of students' evaluation of teaching effectiveness: cognitive and affective criteria. *J. Educ. Psychol.*, 72, 468–475, doi: 10.1037/0022-0663.72.4.468

*Mazor, K., Clauser, B., Cohen, A., Alper, E., and Pugnaire, M. (1999). The dependability of students' ratings of preceptors. *Acad. Med.*, 74, S19–S21, doi: 10.1097/00001888-199910000-00028

McNamara, T. F., and Knoch, U. (2012). The Rasch wars: the emergence of Rasch measurement in language testing. *Lang. Test.* 29, 555–576. doi: 10.1177/0265532211430367

Mushquash, C., and O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analysis. *Behav. Res. Methods* 38, 542–547. doi: 10.3758/BF03192810

Onwuegbuzie, A. J., Daniel, L. G., and Collins, K. M. T. (2006). A meta-validation model for assessing the score-validity of student teaching evaluations. *Qual. Quant.* 43, 197–209. doi: 10.1007/s11135-007-9112-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. *J. Clin. Epidemiol.* 134, 103–112. doi: 10.1016/j.jclinepi.2021.02.003

Park, B., and Cho, J. (2023). How does grade inflation affect student evaluation of teaching? *Assess. Eval. High. Educ.* 48, 723–735. doi: 10.1080/02602938.2022.2126429

*Quansah, F. (2020). An assessment of lecturers' teaching using generalisability theory: a case study of a selected university in Ghana. *South Afr. J. High. Educ.*, 34, 136–150. doi: 10.20853/34-5-4212

Quansah, F. (2022). Item and rater variabilities in students' evaluation of teaching in a university in Ghana: application of many-facet Rasch model. *Heliyon* 8, e12548–e12549. doi: 10.1016/j.heliyon.2022.e12548

Rantanen, P. (2013). The number of feedbacks needed for reliable evaluation: a multilevel analysis of the reliability, stability and generalisability of students' evaluation of teaching. *Assess. Eval. High. Educ.* 38, 224–239. doi: 10.1080/02602938.2011.625471

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press.

Raza, S. A., and Khawaja, F. N. (2013). Faculty development needs as perceived by departmental heads, teachers, and students of Pakistani universities. *Lit. Inform. Comput. Educ. J.* 4, 992–998. doi: 10.20533/licej.2040.2589.2013.0132

Raza, S. A., Majid, Z., and Zia, A. (2010). Perceptions of Pakistani university students about roles of academics engaged in imparting development skills: implications for faculty development. *Bull. Educ. Res.* 32, 75–91.

*Rindermann, H., and Schofield, N. (2001). Generalizability of multidimensional student ratings of university instruction across courses and teachers. *Res. High. Educ.*, 42, 377–399, doi: 10.1023/A:1011050724796

Rosli, M. K., Mistima, S., and Rosli, N. (2017). Students' attitude and anxiety towards statistics a descriptive analysis. *Res. Educ. Psychol.* 1, 47–56.

*Samian, Y., and Noor, N. M. (2012). Students' perception of good lecturer based on lecturer performance assessment. *Procedia Soc. Behav. Sci.*, 56, 783–790, doi: 10.1016/j.sbspro.2012.09.716

Shavelson, R. J., and Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: SAGE.

Spooren, P., Brockx, B., and Mortelmans, D. (2013). On the validity of student evaluation of teaching: the state of the art. *Rev. Educ. Res.* 83, 598–642. doi: 10.3102/0034654313496870

*Spooren, P., Mortelmans, D., and Christiaens, W. (2014). Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Stud. Educ. Eval.*, 43, 88–94, doi: 10.1016/j.stueduc.2014.03.001

Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: a theoretical and empirical analysis. *Basic Appl. Soc. Psychol.* 42, 276–294. doi: 10.1080/01973533.2020.1756817

Taut, S., and Rakoczy, K. (2016). Observing instructional quality in the context of school evaluation. *Learn. Instr.* 46, 45–60. doi: 10.1016/j.learninstruc.2016.08.003

Teker, G. T., Güler, N., and Uyanik, G. K. (2015). Comparing the effectiveness of SPSS and EduG using different designs for generalizability theory. *Educ. Sci.: Theory Pract.* 15, 635–645. doi: 10.12738/estp.2015.3.2278

*Üstünlüoğlu, E., and Can, S. (2012). Student evaluation of teachers: a case study of tertiary level. *Int. J. New Trends Educ. Implicat.*, 3, 92–99

*VanLeeuwen, D. M., Dormody, T. J., and Seevers, B. S. (1999). Assessing the reliability of student evaluation of teaching (SET) with generalizability theory. *J. Agric. Educ.*, 40, 1–9, doi: 10.5032/jae.1999.04001

Vlăsceanu, L., Grünberg, L., and Pârlea, D. (2004). *Quality assurance and accreditation: a glossary of basic terms and definitions*. Bucharest: United Nations Educational, Scientific and Cultural Organization.

Wachtel, H. T. (1998). Student evaluation of college teaching effectiveness: a brief review. *Assess. Eval. High. Educ.* 23, 191–212. doi: 10.1080/0260293980230207

Wilhelm, W. B. (2004). The relative influence of published teaching evaluations and other instructor attributes on course choice. *J. Mark. Educ.* 26, 17–30. doi: 10.1177/0273475303258276

Yang, F. M., and Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Arch. Psychiatry* 26, 171–177. doi: 10.3969/j.issn.1002-0829.2014.03.010