European Journal of Education and Pedagogy
Vol 6 | Issue 1 | January 2025
ISSN 2736-4534

RESEARCH ARTICLE

Student Evaluations: Use, Perceptions, and a Figure of Merit for Non-Grade Biased Evaluation

Mahmoud Al Ahmad [1], Lillian Olule [1], Qurban Memon [1,*], and Adel Najar [2]

ABSTRACT

Student Evaluation of Teaching (SET) is a widely spread and widely acknowledged means of evaluating teaching effectiveness in higher education institutions (HEIs). Student evaluation of teaching (SET) has been extensively researched, but there are still contradicting attitudes toward its use and validity, caused mainly by student grade bias. A performance metric is needed that can be used to identify student evaluation ratings independent of grading bias. This paper explores various purposes for which SET can be utilized, and the perceptions of the key stakeholders on its relative importance. To remove grade bias from evaluation, a figure of merit (F.O.M) is derived that can be used to identify a student evaluation rating independent of grading bias. The F.O.M is derived using statistical analysis that considers preprocessing of data, and removal of outliers to make it robust. It is found that F.O.M values between 0.7 and 1.3 reflect non-grade bias. Based on this F.O.M, recommendations are built to support the validity of SET in the whole process of teaching effectiveness alongside increased stakeholder engagement.

Keywords: Assessment, student evaluation, teaching effectiveness, teaching evaluation.

Submitted: November 19, 2024

Published: January 23, 2025

10.24018/ejedu.2025.6.1.902

[1] *Department of Electrical Engineering, UAE University, United Arab Emirates.*
[2] *Department of Physics, UAE University, United Arab Emirates.*

* ***Corresponding Author:***
e-mail: qurban.memon@uaeu.ac.ae

1. Introduction

Student evaluation of teaching (SET) in higher education institutions (HEIs) has become a key mechanism by which teaching effectiveness is measured. Its origins can be traced back to the first teacher rating scale published in 1915 as reported by Spencer and Flyr (1992). Gupta *et al*. (2018) explore how gender and socioeconomic disparities influence students' evaluations of their teachers in higher education. To evaluate the effectiveness of socio-economic policies, it examines whether these demographic factors shape students' perceptions of their teachers. By applying multivariate and univariate general linear models to analyze student ratings based on gender and socioeconomic factors respectively, the research reveals that socio-economic status, gender-conforming, and non-conforming behaviors, as well as same-gender and cross-gender biases, all contribute to variations in teacher ratings across different disciplines. Xu *et al*. (2018) explore the integration of advanced teaching technologies in the development of smart campuses. A crucial element of smart campuses is the evaluation of teaching performance, which utilizes data on teaching and teacher-student interactions to improve teaching quality and strengthen institutional competitiveness. The paper introduces a smart campus architecture model and a framework for an education data collection and storage platform, serving as a guide for smart campus construction.

Price *et al*. (2017) examined student evaluations of teaching quality at a Swedish university, utilizing data from the Course Experience Questionnaire. Over a decade, ratings were gathered from male and female students. The findings showed some variations, with teachers receiving higher ratings in subjects that were less typical for their gender. Differences in ratings were also noticed based on the gender of both teachers and students, but no consistent bias was found in students favoring teachers of the same gender. Although some effects were statistically significant, they were minor and unlikely to have practical significance. Buragohain *et al*. (2024) examined the effects of immersive learning, where teachers engage in realistic scenarios to hone their skills and receive feedback in a controlled, safe setting, on their overall effectiveness. Immersive learning improves teachers' subject knowledge, teaching skills, and confidence in handling real-world classroom situations. The study highlights the need for additional research, particularly on the long-term impacts of immersive learning,

and underscores its potential to enhance teacher education and professional development initiatives.

To enhance classroom teaching quality and assess its effectiveness, a multimedia teaching quality evaluation system was developed for colleges and universities (Jian, 2019). Using an algorithm, the system analyzed the state of multimedia teaching and the evaluation process through surveys and teacher interviews. The system evaluates courseware, the teaching process, and teaching effects, and manages information related to teachers, students, and evaluations. The results demonstrated that the system effectively evaluates teaching quality, making it a valuable tool for improving classroom instruction. Tseng *et al*. (2018) present a system for analyzing textual opinions from teaching evaluation questionnaires, providing useful reference materials. These questionnaires are treated as educational data and analyzed using data mining techniques. The research employs text sentiment analysis to quantify students' opinions and generate feedback on teaching staff. The findings suggest that classifiers that consider time series factors produce more precise sentiment analysis.

Okoye *et al*. (2023) investigate the role of Technology-Enhanced Learning (TEL) in higher education, with a focus on the difficulties educators encounter in selecting appropriate tools, methods, and technologies to maintain ongoing learning. The findings suggest that adaptable digital frameworks and creative teaching strategies can effectively promote continuous learning and contribute to pedagogical advancements. To meet the demand for intelligent educational evaluation systems due to the widespread adoption of online education, Pei and Lu (2023) introduce a deep-learning-based evaluation system. By utilizing the Offset Minimal Sum (OMS) algorithm, the network reduces parameters by approximately 59.64% and cuts training time by 54.92% compared to other models. Experimental results also show enhanced performance in classifying unbalanced data. The contemporary state-of-art research on student evaluations of teaching (SET) encompasses several key areas:

- Validity: Evaluates whether SET tools reliably measure teaching effectiveness across different contexts.
- Bias: Examines gender, racial, and other biases in SET, and seeks ways to mitigate them to ensure fairness.
- Feedback mechanisms: Examines alternative methods like peer evaluations and qualitative feedback to provide more detailed insights into teaching effectiveness beyond traditional surveys.
- Impact on Teaching Practices: Investigates how SET results affect instructors' teaching methods and professional development to improve their effectiveness.
- Student Perceptions: Analyzes how students perceive SET results, as well as their attitudes toward feedback.
- Ethical Considerations as guidelines for responsible conduct and interpretation.

Fig. 1 shows how these different facets of SET interact with each other showing a complex interplay.

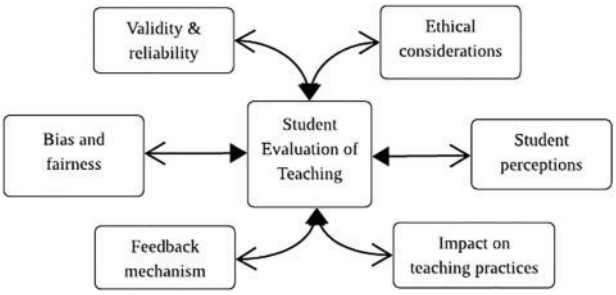

Fig. 1. Interaction among different facets of SET.

The study focuses on these facets and investigates the use and perceptions of student evaluations. However, a significant challenge in SET is the presence of biases. This research aims to address bias within SET and thus increase the awareness among stakeholders about the importance of SET by building a performance metric in a statistical way to reflect non-grade bias student evaluation. The extended role of SET beyond a personal tool for faculty to improve teaching effectiveness is presented.

The research question addressed in this study focuses on identifying and mitigating specific sources of bias, such as those related to teacher demographics, student preconceptions, and external factors. It aims to incorporate a statistical framework to significantly reduce demographic-related biases (e.g., age, gender, ethnicity) in student evaluations of teaching, leading to more accurate and objective assessments of teaching effectiveness.

2. Methods and Impact

Alternative methods to traditional SET offer institutions more meaningful feedback, fostering continuous improvement in teaching practices (Memon, 2007; Memon & Harb, 2009; Memon & Khoja, 2009). These methods are peer evaluation, self-assessment, student focus groups, teaching portfolios, learning outcomes assessment, and formative feedback. Though the selection of an evaluation method hinges on institutional priorities, available resources, and the desired depth of feedback, to gain insight into differences, SET may be compared to alternative approaches based on factors such as:

- Feedback Depth: SET provides standardized, quantitative feedback, but lacks depth compared to qualitative methods like peer evaluation, focus groups, and teaching portfolios.
- Validity and Reliability: SET validity and reliability vary due to factors such as survey design and biases. Alternative methods such as peer evaluation and learning outcomes assessment may offer more objective and context-specific feedback.
- Feasibility and Resources: SET is easy to administer, but data analysis can be time-consuming. Alternatives like peer evaluation and teaching portfolios require more resources and training.
- Impact on Teaching Improvement: While SET identifies improvement areas but may have limited impact, peer evaluation and formative feedback

offer more actionable insights, leading to targeted improvements.
- Ethical Considerations: SET raises ethical concerns like anonymity and biases. Alternative methods may pose similar concerns, such as ensuring confidentiality in peer evaluations.

2.1. Students as Evaluators

The debate over students as evaluators has both supporters and critics. Proponents argue that students, as the primary "end-users" who spend the most time with faculty, are in the best position to assess teaching effectiveness. Opponents, however, claim that students have a limited understanding of effective teaching and may not accurately judge factors like course comprehensiveness or long-term objectives (Nasser & Barbara, 2002; Sojka *et al*., 2002). Despite these concerns, it is generally accepted that with well-designed evaluation tools, students can provide reliable and valuable feedback.

Student evaluations can be conducted using quantitative instruments like questionnaires and surveys (Malgorzata *et al*., 2016; Richardson, 2005), or qualitative methods such as dialogue-based evaluations (Rebecca & Dobbins, 2013; Steyn *et al*., 2019). While there is less research on qualitative tools (Borch *et al*., 2020), they are believed to offer more detailed, context-specific feedback useful for course improvements. However, qualitative assessments are more labor-intensive, often requiring multiple tools (e.g., interviews, observations, and documents) to provide a holistic view of student learning (Gardner *et al*., 2012). This makes the evaluation process more time-consuming, particularly in data collection and analysis.

In higher education institutions (HEIs), the most commonly used evaluation tools are online or paper-based questionnaires, typically administered mid-semester or at the end. While both methods yield similar evaluation scores, online surveys generally have lower response rates (Gardner *et al*., 2012). To address this, strategies such as offering grade incentives (Dommeyer *et al*., 2004), sending repeated email reminders, and offering prizes have been proposed (Nulty, 2008). Research shows that the more methods employed to increase participation, the higher the response rate for online surveys (Ballantyne, 2005; Ogier, 2005).

2.2. Impact and Use of Student Evaluations

Originally, the purpose of student evaluations was to enhance teaching effectiveness and student learning (Leckey & Neville, 2001). However, over time, other uses emerged. MacLeod (2000) offers a faculty-centered classification, identifying two main functions: staff appraisal and staff development. Additionally, the evaluation system's design tends to emphasize features supporting its intended purpose. Sproule (2000) proposed a broader classification, considering the institution, faculty, and students, and categorized the uses into two primary functions: summative (for evaluation) and formative (for improvement).

The summative function includes all functions concerning employment. It provides a quick, low-cost means of generating a general view of teaching and course delivery (Surgenor, 2013) of staff, primarily for administrative decisions. The major uses that fall under this summative function include:

- Promotion/tenure decisions and salary raise (Dommeyer *et al*., 2004; Nulty, 2008; Wattiaux *et al*., 2010): Many higher education institutions use student evaluations as a key metric in promotion, tenure decisions, and salary raises, alongside research and service (McCabe & Layne, 2012). Administrators rely on these evaluations to assess faculty performance for tenure or promotion. Wattiaux *et al*. (2010) suggest improving this process by ensuring reliable evaluation tools, updating guidelines for faculty on evaluation usage in promotion/tenure, and providing mentorship. However, Stephen and McKelvey (2019) warn that policy language must be clear for such evaluations to be valid in these decisions.
- Faculty merit and Awards (Ballantyne, 2005): Student ratings are also used to recognize and reward faculty. For instance, Harvard uses evaluations to award lecturers, teaching assistants, and fellows the Certificate of Distinction in Teaching (Gravestock & Gregor-Greenleaf, 2008). Similarly, the University of Toronto includes evaluations as part of the criteria for the President's Teaching Award (Gravestock & Gregor-Greenleaf, 2008). Such recognitions boost faculty morale and reduce staff turnover (Kalis & Kirschenbaum, 2008).
- Student evaluations have recently been incorporated into the documentation required for some faculty positions (Madichie, 2011). While reference letters traditionally offer professional and character assessments from the institution's perspective, student evaluations provide feedback from the student's viewpoint.
- Student evaluations provide valuable insights to help students choose courses and faculty, especially for electives (Babad, 2001). Babad's study on student decision-making found that the first elective choice was primarily influenced by the lecturer's quality and the course's learning value, while the second choice was more influenced by how easy and comfortable the course appeared to be.

Sproule defined the formative function as the function that relates to aspects concerning quality enhancement. The formative uses of student evaluation have been universally accepted as an important tool for academic development. They provide a continuous communication channel between faculty and students and establish a system for the improvement of student learning experiences. The key formative uses of SET include:

- Feedback for Improving Teaching Practices: SET helps faculty identify strengths and weaknesses in their teaching. Combined with other feedback, such as peer evaluations (Lomas & Nicholls, 2005), it provides a comprehensive perspective for continuous improvement (Chan *et al*., 2014).

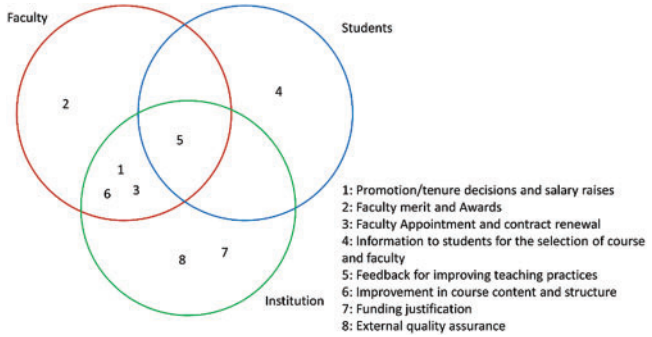

Fig. 2. Functions of SET and their relative importance.

- Improvement in course content and structure: SET offers feedback on aspects like lecture format, materials, and assessment methods, helping faculty refine course design. This feedback also supports program reviews, guiding decisions on course revisions or discontinuation to improve educational offerings (Scheepers, 2019).
- Funding justification: SET plays a role in justifying funding by assessing faculty quality measured by indicators including availability to students as determined by student evaluations (Schmidt, 1999).
- External quality assurance: External quality assurance agencies require HEIs to prove the quality of their teaching (Barrie, 2000) for purposes of accreditation (Martin & Gibbs, 2001). Student evaluations provide supportive evidence to validate that the threshold of quality has been achieved.

Fig. 2 summarizes the perceived relative importance of the functions of SET from the perspectives of the three key stakeholders. Evidently, SET is viewed as more important to faculty/instructors than to students. This is unsurprising as out of the eight stated purposes of SET, only three could be considered by students as having a direct impact on them.

Herbert and Roche (1997) identified nine dimensions of teaching effectiveness: workload/difficulty, assignments, examination grading, course coverage, individual rapport, group interaction, organization, enthusiasm, and learning value (Clayson & Sheffet, 2006). Since including all these dimensions in a single student evaluation can be challenging, many higher education institutions use additional assessment methods for a more comprehensive evaluation. Peer reviews, in-classroom observations, and teaching portfolios are commonly used. Teaching portfolios, in particular, are widely used and include elements such as teaching philosophy statements, course materials, sample student work, evidence of teaching awards, professional development, and research on teaching and learning (Gravestock & Gregor-Greenleaf, 2008).

2.3. *Stakeholder Perceptions*

It is reported that stakeholders have both positive and negative attitudes toward the purpose, validity usefulness, and consequences of SET (Surgenor, 2013). To a great extent, most researchers believe that student evaluations are valid, reliable, and useful owing to the decades of research done on the subject (Gupta *et al*., 2018). Faculty perceived that SET ratings are a rating of personality or character (Herbert & Roche, 1997), with more popular faculty getting higher ratings than less popular faculty. It was perceived that these ratings do not take into account other factors that contribute to teaching effectiveness such as knowledge of subject matter, technical competence, and use of innovative teaching technologies (Aziz & Islam, 2022). Indeed, this does reflect a difference in opinion as to what faculty and students believe constitutes good teaching.

Faculty often believe that Student Evaluations of Teaching (SET) influence course content (Sojka *et al*., 2002) and teaching styles, while students are less convinced, possibly because SETs are usually administered at the end of a course, with changes occurring in subsequent terms. This delay means students may not see the impact of their feedback. To address this, researchers suggest implementing mid-semester evaluations, which would allow faculty to make adjustments during the course and enable students to observe the effects of their feedback. One university developed a web-based form for anonymous, real-time feedback, allowing students to provide input after difficult assignments or with concerns, that instructors could address promptly (Sojka *et al*., 2002).

Basow *et al*. (2013) found that both gender and race impact student evaluations. Male students generally rated instructors higher than female students did, and male professors received better evaluations than female professors. These disparities may stem from gender stereotypes, where women are expected to be more nurturing and are judged more critically. Minority instructors also received lower ratings (Foschi, 2000), which can be attributed to stereotypes that suggest they are less competent than majority group members. Consequently, minority instructors often feel they must work harder to be seen as equally competent and perceive that mistakes are judged more harshly (Churchill, 2006).

A common perception on the part of faculty is the positive correlation between the grade and SET rating (Nasser & Barbara, 2002). The negative side effect of this influence is that staff may be inclined to grade students more leniently (Churchill, 2006) to achieve a higher evaluation score. This may be particularly true for non-tenured faculty aiming for a tenured position.

3. Results

Research shows mixed results regarding the correlation between student evaluations and grades. Some studies find no significant relationship (Marlin & Gaynor, 1989), while others report a positive correlation, generally ranging from 0.2 to 0.5 (Kulik, 2001). Kulik generalized that typically these positive correlation coefficients fall in the range of between 0.2 to 0.3 (Linse, 2017). A survey by Linse (2017) identified a wider range of 0.1 to 0.5 for evaluations conducted in a laboratory setting. Positive correlation can be interpreted in at least six different ways highlighted in the literature. One interpretation is that better teachers attract better students or quality teachers teach more effectively, resulting in better student learning and ultimately higher grades. Another interpretation is that students tend to give

higher ratings to faculty who give higher grades (Feldman, 1976). It is therefore seen that by inflating grades faculty can influence their rating. The positive correlation between grades and evaluation may also be attributed to the student's adeptness at doing work. Thus, higher grades are causally dependent on the student's academic ability. It is also posited that more academically adept students are more capable of discerning good teaching and therefore give more favorable evaluations than other students. Costin *et al*. (1971) argue that a positive relationship between grades and evaluation may be fully or partly attributed same antecedent variable, the student's interest or motivation for the course. Research has shown students taking a course as an elective show a more favorable evaluation than those taking it as a requirement (Costin *et al*., 1971). Student interest in a course can often stem from the instructor's ability to engage and stimulate interest, which reflects their teaching effectiveness. Additionally, various factors such as the instructor's rank, age, gender, class size, course level, and difficulty can also impact both grades and student evaluations, either directly or indirectly (Costin *et al*., 1971).

Many interpretations of student evaluations emphasize the main effects and overlook interactive factors. For example, the link between grades and evaluations may be affected by students' expectations of their future grades. Instructors who receive high evaluations might have students who later receive lower grades. Additionally, students often misperceive their grades—poorer students tend to overestimate them, while better students may underestimate them.

Typically, developing a figure of merit for SET involves thoughtful consideration of institutional goals, and context, alongside best practices in assessment and evaluation. Here are the steps to develop a figure of merit for SET:

- Define Goals and Objectives: Clearly articulate the goals and objectives of the SET process, focusing on essential aspects of teaching effectiveness and desired outcomes.
- Identify Key Metrics: Determine the metrics or indicators needed to evaluate teaching effectiveness, encompassing both quantitative measures and qualitative feedback.
- Select Evaluation Instruments: Choose or design suitable evaluation tools aligned with SET objectives and encompassing relevant facets of teaching effectiveness.
- Establish Scoring Criteria: Develop objective scoring criteria for each metric to ensure fairness in assessment.
- Collect Data: Administer evaluations anonymously after each term or course to gather student feedback.
- Analyze Results: Analyze the data to compute metrics, consolidate ratings, and review qualitative feedback.
- Interpret Findings: Interpret SET outcomes about predefined goals, identifying strengths, areas for improvement, and emerging patterns in teaching effectiveness.
- Utilize Feedback for Improvement: Utilize SET feedback to drive faculty development initiatives, provide support for instructors in refining their teaching methods, and enhance overall teaching effectiveness.
- Monitor and Review: Continuously monitor and evaluate the SET process, making necessary adjustments based on feedback from stakeholders and institutional priorities.

By adhering to these steps, institutions can devise a figure of merit for SET that effectively evaluates teaching effectiveness and promotes ongoing enhancement in teaching practices. Below, a similar approach is adopted to develop a figure of merit for SET that eliminates the influence of differential grading standards:

- As a pre-processing step, the student rating and student grade should be adjusted to the same scale. For example, in Tables I–IV 4, the original rating data is out of 20 while the original grade data is out of 100. All sets of data are rescaled to be out of 4.
- After students have submitted their instructor evaluation, the collected responses are processed by finding the corresponding average rating ($\overline{x}$). This average rating will be used to remove outliers.
- Outliers are identified as any values greater than ±30% of the average rating. This percentage is based on statistical analysis. Also from the literature, 12% to 25% of the variance in average rating can be attributed to grading (Feldman, 1976), and therefore its impact is most significant where unusually high or low grades are given.
- From the new dataset of collected responses with outliers removed, the new average rating ($\overline{xn}$) is determined.
- The same procedure is applied to the student grades. The average grade ($\overline{y}$) is found, outliers beyond ±30% of the average grade are removed and the new average grade ($\overline{yn}$) is identified.
- The figure of merit (F.O.M) is determined from the ratio of the new student rating average to the new grade average.

$$F.O.M = \overline{xn}/\overline{yn} \tag{1}$$

The F.O.M can be related to the teaching effectiveness as follows; mean ratios of one or larger are indicative of greater teaching effectiveness and mean ratios less than one indicate the converse. This would suggest that an instructor who gives good grades with a lower evaluation is a poor teacher. To put it another way, it would suggest that an instructor who gave low grades and still got high evaluations was a good teacher.

The F.O.M represents a simple rating/grade ratio for control for biases associations, instead of a ratio of student grades by normative grades held by the institution or in the perception of the students. The algorithm does not use normalized values, it uses the ratio after removing outliers, maintaining the variance and skewness of the variables. Using normalized values would imply that the mean ($\overline{xn}$) would have the same value as ($\overline{yn}$) and the ratio would always be equal to one.

TABLE I: F.O.M from 2018 Student Rating and Student Grade Data

	Rating (out of 20)	Grade (out of 100)	Rating (out of 4)	Grade (out of 4)	Refined rating*	Refined grade#	FOM
1	12.85	34.75	2.57	1.39	2.57	1.39	1.43
2	16.65	80.00	3.33	3.20			
3	14.75	82.50	2.95	3.30			
4	6.90	19.25	1.38	0.77			
5	11.35	62.75	2.27	2.51	2.27		
6	7.60	39.25	1.52	1.57	1.52	1.57	
7	3.65	19.25	0.73	0.77			
		Average	2.11	1.93	2.12	1.48	

Note. *Only ratings within ±30% of the average rating (2.11), and average grade (1.93) were considered.

TABLE II: F.O.M from 2019 Student Rating and Student Grade Data

	Rating (out of 20)	Grade (out of 100)	Rating (out of 4)	Grade (out of 4)	Refined rating*	Refined grade*	FOM
1	14.15	60.00	2.83	2.40	2.83	2.40	1.03
2	18.05	84.25	3.61	3.37			
3	16.95	75.50	3.39	3.02			
4	6.63	39.90	1.33	1.60		1.60	
5	10.83	59.50	2.17	2.38	2.17	2.38	
6	8.00	35.50	1.60	1.42	1.60		
7	3.38	18.00	0.68	0.72			
		Average	2.23	2.13	2.20	2.13	

Note. *Only ratings within ±30% of the average rating (2.23), and average grade (2.13) were considered.

To validate the methodology, it was applied to three sets of student evaluations and corresponding student grade data taken from three subsequent years of a course intake with a class size of n = 7 and a survey response rate of 100%. The details and calculations are shown in Tables I–III for 2018, 2019, and 2020 data, respectively. Another student grading and rating data for a class size of 18 with a 95% response rate was also used to derive F.O.M., and the results are shown in Table IV.

The results of F.O.M. of 1.43, 1.03, and 1.24 calculated for the years 2018, 2019, and 2020, respectively, show that over the three years, the instructors' teaching was considered effective by all three cohorts of students since F.O.M. was greater than one. The highest rating was achieved in 2018, and the lowest rating was achieved in 2019. The teaching methods used in these years can be compared to identify methods or techniques to improve student learning. This could also possibly be the reason why the rating improved in 2020.

Calculations of the correlation coefficients for the three years reveal a strong correlation between the student grades and student ratings with r = 0.87, r = 0.97, and r = 0.90 ($\rho < 0.05$ in all cases). Excluding other assessment methods, it may be challenging to determine if this is an accurate reflection of teaching quality or the influence of grading. It should be noted that the variance indicated by a correlation simply indicates the amount of variance in one variable that can be accounted for by another variable. That variance is over the entire range of the variable. The correlation is not related to the cut-off point at ± 30% which is justified by r2 measures. In addition, extreme scores are just as likely to result from other associative variables as they are from the grade vs. evaluation association.

We further propose that for consideration of the student rating for the purposes discussed in the earlier section, the F.O.M should lie within the range of ±30% of the nominal value. Alternatively, the F.O.M should lie in the range of 0.7 to 1.3. F.O.Ms outside of this range may be indicative of an anomaly. It should be noted that the method was applied to small-size classes (with three cohorts), it can be applied to a relatively large population with similar results. To illustrate the proposed approach, the figure-of-merit was applied to both a smaller-size elective course and a relatively larger-size required course. The elective course was chosen (with

TABLE III: F.O.M from 2020 Student Rating and Student Grade Data

	Rating (out of 20)	Grade (out of 100)	Rating (out of 4)	Grade (out of 4)	Refined rating*	Refined grade#	FOM
1	13.67	43.80	2.73	1.75	2.73	1.75	1.24
2	18.52	89.75	3.70	3.59			
3	16.86	88.35	3.37	3.53			
4	7.48	18.80	1.50	0.75			
5	9.81	60.50	1.96	2.42	1.96	2.42	
6	7.48	37.85	1.50	1.51		1.51	
7	3.12	19.05	0.62	0.76			
		Average	2.20	2.05	2.35	1.90	

Note. *Only ratings within ±30% of the average rating (2.20), and average grade (2.05) were considered.

TABLE IV: F.O.M FROM 2020 STUDENT RATING AND STUDENT GRADE DATA

	Rating (out of 20)	Grade (out of 100)	Rating (out of 4)	Grade (out of 4)	Refined rating*	Refined grade#	FOM
1	14.67	45.80	2.93	1.83	2.93		0.97
2	18.52	89.75	3.70	3.59		3.59	
3	16.86	88.35	3.37	3.53	3.37	3.53	
4	8.48	68.80	1.70	2.75		2.75	
5	9.81	70.50	1.96	2.82	1.96	2.82	
6	7.48	67.85	1.50	2.71		2.71	
7	9.12	69.05	1.82	2.76		2.76	
8	6.52	46.75	1.30	1.87			
9	14.52	72.66	2.90	2.91	2.90	2.91	
10	16.32	82.32	3.26	3.29	3.26	3.29	
11	11.56	65.32	2.31	2.61	2.31	2.61	
12	16.82	86.55	3.36	3.46	3.36	3.46	
13	17.32	78.32	3.46	3.13		3.13	
14	18.30	82.56	3.66	3.30		3.30	
15	15.31	70.61	3.06	2.82	3.06	2.82	
16	14.45	45.52	2.89	1.82	2.89		
17	8.66	52.66	1.73	2.11		2.11	
		Average	2.64	2.78	2.89	2.99	

Note. *Only ratings within ±30% of the average rating (2.64), and average grade (2.78) were considered.

a limited class size of 7 in the years 2018, 2019, and 2020), and a required course (with a class size of 22 in the year 2020). As expected, in conformity with (Adams *et al*., 2021; Wang & Calvano, 2022), the results are better with smaller class sizes than relatively higher ones.

The validity and reliability of student evaluations of teaching (SET) are essential for trustworthy feedback. Key factors are validity, reliability, bias, and sample size. The validity includes content, construct, face, and criterion. The reliability includes consistency over time. These factors help ensure SET questionnaires evaluate teaching effectively. The SET questionnaires used in this study have been developed and regulated by accreditation agencies for the last 20 years. The results obtained in three SET calculations (with a 100% response rate in a small class size and a 95% response rate in another relatively large class size) are statistical and thus can be claimed without bias. Thus, the results obtained are valid and reliable. However, it is felt that additional results are needed with higher class sizes to generalize the result.

4. DISCUSSION

This study focused on how SET can be made independent of course grade bias, and thus be useful in overall teaching evaluation. For this purpose, the relation between student grades and rating data was examined. It was found that the F.O.M. parameter should be used alongside student learning outcomes and peer evaluations. Considering the essential nature of SET for teaching effectiveness, the following recommendations can be made:

- Stakeholders are often the least aware of the usefulness and impact of SET. Evaluation instruments could include details on their purpose or highlight past improvements based on SET results to improve awareness. If this could introduce bias, alternative communication methods, such as emails, may be used instead.
- As education evolves, institutions must regularly review the validity and relevance of SET instruments. For instance, with the shift to online learning, traditional SET questionnaires may not capture the key factors affecting teaching effectiveness. New evaluations should reflect these changes in learning environments.
- Student evaluations typically occur at the end of the semester, but more frequent evaluations could be beneficial by capturing events that students might forget later. Regular evaluations would help faculty identify effective teaching methods earlier. The figure of merit (F.O.M.) from both mid-semester and end-of-semester evaluations could track instructor improvements or averaging the two could provide a more objective rating.
- The shift toward online learning presents challenges, such as lower student response rates (Wang & Calvano, 2022) for evaluations. To improve participation without introducing bias, institutions could offer incentives. One suggestion is a grade incentive awarded only if students complete evaluations for all their courses, preventing selective evaluation of courses where they anticipate lower grades.
- Student evaluations used for summative purposes must account for bias to ensure fairness, as they often reflect student perceptions influenced by bias. Research by Adams *et al*. (2021) revealed significant gender bias, especially against women, which can lead to inaccurate assessments and affect faculty performance reviews or hiring decisions, potentially harming long-term career prospects. To address this, institutions could implement a "bias correction factor" tailored to their specific context in faculty evaluation ratings.
- SET ratings are considered valid only if the response rate reaches at least 30%, with a 3%

sampling error and 95% confidence level, as recommended by Nulty (2008). This ensures results represent the overall class experience rather than individual views, reducing non-response bias. The response rate reflects the percentage of usable responses from total enrolments, with all exclusions explained. Analyzing the response rate for each evaluation item can also help identify specific issues.

- Lastly, we recommend the use of the proposed methodology to determine student ratings in the range of 0.7 to 1.3 being independent of grade bias, as ratings outside this range are considered anomalous.

5. Conclusion

The role of SET has expanded beyond its original purpose as a tool for faculty improvement. An F.O.M. was developed to evaluate teaching quality, using normalized student grades and ratings within ±30% of the averages to reduce grade bias. F.O.M values above 0.7 indicate good teaching, while values below 0.7 suggest poorer quality. An F.O.M range of 0.7 to 1.3 is recommended for both formative and summative evaluations, as it reflects less biased assessments. The F.O.M thus calculated is more robust than plain rating data, providing a practical way to minimize grade influence without overly complex methods.

There are no ramifications of this approach since the F.O.M. is derived from grading and rating data and is, thus, independent of any other course grade and respective rating. Each course taught by an instructor will have its own F.O.M. as representative value. This new statistical value of SET can be graphed on a timeline to visualize how each program course was taught by the same or by different instructors.

Reducing bias in student evaluations of teaching (SET) has several practical implications that can significantly improve both educational practices and institutional decision-making. The key areas are:

- Fairer Teacher Assessments lead to more equitable tenure and foster a more diverse faculty.
- Enhanced Teaching Quality enables more accurate feedback on instructional methods and student engagement and promotes targeted professional development.
- Improved Student Learning Outcomes help focus on improving teaching and more effective course delivery.
- Institutional Reputation and Accountability for fairness, improved legal compliance, and reduced grievances.
- Increased Faculty Retention and Morale leads to higher retention and reduced burnout.

Conflict of Interest

The authors declare that they do not have any conflict of interest.

References

Adams, S., Bekker, S., Fan, Y., Gordon, T., Shepherd, L., Slavich, E., & Waters, D. (2021). Gender bias in student evaluations of teaching: 'Punish[ing] those who fail to do their gender right. *Higher Education*, *83*, 787–807.

Aziz, S., & Islam, S. (2022). Impact of mixed pedagogy on engineering education. *IEEE Transactions on Education*, *65*(1), 56–63.

Babad, E. (2001). Students' course selection: Differential considerations for first and last course. *Research in Higher Education*, *42*(4), 469–492.

Ballantyne, C. (2005). Moving student evaluation of teaching online: Reporting pilot outcomes and issues with a focus on how to increase student response rate. *Australasian Evaluations Forum: University Learning and Teaching: Evaluating and Enhancing the Experience*.

Barrie, S. (2000). Reflections on student evaluation of teaching: Alignment and congruence in a changing context. *Refereed Proceedings of Teaching Evaluation Forum*.

Basow, S., Codos, S., & Martin, J. (2013). The effects of professors' race and gender on student evaluations and performance. *College Student Journal*, *47*(2), 352–363.

Borch, I., Sandvoll, R., & Risør, T. (2020). Discrepancies in purposes of student course evaluations: What does it mean to be 'satisfied'? Educational Assessment. *Evaluation and Accountability*, *32*(1), 83–102.

Buragohain, D., Deng, C., Sharma, A., & Chaudhary, S. (2024). The impact of immersive learning on teacher effectiveness: A systematic study. *IEEE Access*, *12*, 35924–35933.

Chan, C., Luk, L., & Zeng, M. (2014). Teachers' perceptions of student evaluations of teaching. *Educational Research and Evaluation*, *20*(4), 275–289.

Churchill, L. (2006). Professor Goodgrade: Or how I learned to stop worrying and give lots of A's. *Chronicle of Higher Education*, *52*, 25–30.

Clayson, D., & Sheffet, M. (2006). Personality and the student evaluation of teaching. *Journal of Marketing Education*, *28*(2), 149–160.

Costin, F., Greenough, W., & Menges, R. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research*, *41*(5), 511.

Dommeyer, C., Baum, P., Hanna, R., & Chapman, K. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, *29*(5), 611–623.

Feldman, K. (1976). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education*, *4*(1), 69–111.

Foschi, M. (2000). Double standards for competence: Theory and research. *Annual Review of Sociology*, *26*(1), 21–42.

Gardner, M., Hickmott, J., & Ludvik, M. (2012). *Demonstrating Student Success: A Practical Guide to Outcomes-Based Assessment of Learning and Development in Student Affairs*. Stylus Publishing.

Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student Course Evaluations: Research, Models, and Trends*. Higher Education Quality Council of Ontario.

Gupta, A., Garg, D., & Kumar, P. (2018). Analysis of students' ratings of teaching quality to understand the role of gender and socio-economic diversity in higher education. *IEEE Transactions on Education*, *61*(4), 319–327.

Herbert, M., & Roche, L. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52*(11), 1187–1197.

Jian, Q. (2019). Multimedia teaching quality evaluation system in colleges based on genetic algorithm and social computing approach. *IEEE Access*, *7*, 183790–183799.

Kalis, M., & Kirschenbaum, H. (2008). Faculty awards at US colleges and schools of pharmacy. *American Journal of Pharmaceutical Education*, *72*(4), 1–6.

Kulik, J. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research*, *109*(1), 9–26.

Leckey, J., & Neville, N. (2001). Quantifying quality: The importance of student feedback. *Quality in Higher Education*, *7*(1), 19–32.

Linse, A. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation*, *54*, 94–106.

Lomas, L., & Nicholls, G. (2005). Enhancing teaching quality through peer review of teaching. *Quality in Higher Education*, *11*(2), 137–149.

MacLeod, C. (2000). Student feedback on teaching: An institutional perspective. *Refereed Proceedings of Teaching Evaluation Forum*.

Madichie, N. (2011). Students' evaluation of teaching (SET) in higher education: A question of reliability and validity. *The Marketing Review*, *11*(4), 381–391.

Malgorzata, E., Erikson, M., & Punzi, E. (2016). Student responses to a reflexive course evaluation. *Reflective Practice*, *17*(6), 663–675.

Marlin, J., & Gaynor, P. (1989). Do anticipated grades affect student evaluations? A discriminant analysis approach. *College Student Journal*, *23*(2), 184–192.

Martin, C., & Gibbs, G. (2001). The evaluation of the student evaluation of educational quality questionnaire (SEEQ) in UK higher education. *Assessment & Evaluation in Higher Education*, *26*(1), 89–93.

McCabe, K. A., & Layne, L. S. (2012). The role of student evaluations in tenure and promotion: Is effective teaching really being measured. *The Department Chair*, *22*(3), 17–20.

Memon, Q. (2007). On analysis of electrical engineering programme in GCC countries. *European Journal of Engineering Education*, *32*(5), 551–560.

Memon, Q., & Harb, A. (2009). Developing electrical engineering education program assessment process at UAE University. *Australasian Journal of Engineering Education*, *15*(3), 155–164.

Memon, Q., & Khoja, S. (2009). Semantic web approach to academic program assessment. *International Journal of Engineering Education*, *25*(5), 1020–1028.

Nasser, F., & Barbara, F. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education*, *27*(2), 187–198.

Nulty, D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, *33*(3), 301–314.

Ogier, J. (2005). The response rates for online surveys—a hit and miss affair. In *Australasian evaluations*. University Learning and Teaching: Evaluating and Enhancing the Experience. https://www.academia.edu/927234/The_response_rates_for_online_surveys_a_hit_and_miss_affair.

Okoye, K., Daruich, S., De La, O. J., Castaño, R., Escamilla, J., & Hosseini, S. (2023). A text mining and statistical approach for assessment of pedagogical impact of students' evaluation of teaching and learning outcome in education. *IEEE Access*, *11*, 9577–9596.

Pei, Y., & Lu, G. (2023). Design of an intelligent educational evaluation system using deep learning. *IEEE Access*, *11*, 29790–29799.

Price, L., Svensson, I., Borell, J., & Richardson, J. (2017). The role of gender in students' ratings of teaching quality in computer science and environmental engineering. *IEEE Transactions on Education*, *60*(4), 281–287.

Rebecca, F., & Dobbins, K. (2013). Are we serious about enhancing courses? Using the principles of assessment for learning to enhance course evaluation. *Assessment & Evaluation in Higher Education*, *38*(2), 142–151.

Richardson, J. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment & Evaluation in Higher Education*, *30*(4), 387–415.

Scheepers, A. (2019). *SET project: Student evaluations of teaching, measuring and enhancing course quality and teaching quality* [Unpublished master's thesis]. Rotterdam School of Management.

Schmidt, P. (1999, July 2). *A State Transforms Colleges with 'Performance Funding'*. The Chronicle of Higher Education. https://www.chronicle.com/article/a-state-transforms-colleges-with-performance-funding/.

Sojka, J., Gupta, A., & Deeter-Schmelz, D. (2002). Student and faculty perceptions of student evaluations of teaching: A study of similarities and differences. *College Teaching*, *50*(2), 44–49.

Spencer, P., & Flyr, M. (1992). *The Formal Evaluation as an Impetus to Classroom Change: Myth or Reality?* University of California Press.

Sproule, R. (2000). Student evaluation of teaching: Methodological critique. *Education Policy Analysis Archives*, *8*(50), 125–142.

Stephen, P., & McKelvey, S. (2019). Canada: Student evaluations in promotion and tenure. *Discovery*, *5*(5), 5–7. https://stewartmckelvey.com/wp-content/uploads/2019/11/Stewart-McKelvey-Discovery-Issue-5-Fall-2019.pdf.

Steyn, C., Davies, C., & Sambo, A. (2019). Eliciting student feedback for course development: The application of a qualitative course evaluation tool among business research students. *Assessment & Evaluation in Higher Education*, *44*(1), 11–24.

Surgenor, P. (2013). Obstacles and opportunities: Addressing the growing pains of summative student evaluation of teaching. *Assessment & Evaluation in Higher Education*, *38*(3), 363–376.

Tseng, C., Chou, J., & Tsai, Y. (2018). Text mining analysis of teaching evaluation questionnaires for the selection of outstanding teaching faculty members. *IEEE Access*, *6*, 72870–72879.

Wang, L., & Calvano, L. (2022). Class size, student behaviors, and educational outcomes. *Organization Management Journal*, *19*(4), 126–142.

Wattiaux, M., Moore, J., Rastani, R., & Crump, P. (2010). Excellence in teaching for promotion and tenure in animal and dairy sciences at doctoral/research universities: A faculty perspective. *Journal of Dairy Science*, *93*(7), 3365–3376.

Xu, X., Wang, Y., & Yu, S. (2018). Teaching performance evaluation in smart campus. *IEEE Access*, *6*, 77754–77766.