Student Evaluations of Teaching Across Modalities: A Critical Review of Validity, Bias, and Implications for Higher Education

C. Dallas Pierce, MBA

Marywood University

Ph.D. in Strategic Leadership and Administrative Studies

Fall 2025

Abstract

Student evaluations of teaching (SET) are among the most widely used tools for assessing instructional effectiveness in higher education. Despite their ubiquity, research demonstrates that SETs are compromised by validity and reliability concerns, systemic bias against women and minority faculty, and modality-specific inequities that disadvantage online instructors. Furthermore, reliance on SETs as high-stakes measures of performance produces harmful consequences, including faculty stress, incentives for grade inflation, and the erosion of academic rigor. This paper reviews recent literature to synthesize four major themes, validity and reliability, bias, instructional modality, and consequences, before analyzing them through four theoretical frameworks: Institutional Theory, Critical Theory, Role Theory, and Expectancy Theory. The analysis demonstrates that SETs are not neutral measures but structural mechanisms that reproduce inequity, perpetuate institutional inertia, and distort educational priorities. Ethical implications include compromised fairness, systemic discrimination, and threats to both faculty well-being and academic integrity. To address these issues, the paper recommends adopting multi-measure evaluation systems, implementing safeguards against bias, differentiating evaluation frameworks by instructional modality, and enhancing transparency in the use of SET data. The conclusion argues that reforming SET practices is essential to promoting fairness, equity, and quality in higher education, and calls on institutions to adopt comprehensive evaluation policies that align with their stated commitments to inclusion and academic excellence.

Keywords: Student evaluations of teaching (SET), teaching effectiveness, higher education policy, instructional modality, online learning, face-to-face instruction, bias in evaluation, faculty assessment

Table of Contents

Introduction

Student evaluations of teaching (SETs) are widely used across higher education institutions as a primary tool for assessing instructional effectiveness and making critical personnel decisions such as tenure, promotion, and course assignments. The problem is that SET instruments are often applied in the same way to courses delivered in different instructional modalities, specifically online and face-to-face settings, despite evidence that this practice is flawed and unreliable. This issue has been documented across universities globally, including in the United States, the United Kingdom, and other higher education contexts (Daskalopoulou, 2024; Quansah et al., 2024). Over the past five years, a growing body of scholarship has emphasized that the modality of instruction, whether in person, hybrid, or online, affects how students perceive and rate their instructors (Daumiller et al., 2023; Zheng et al., 2023). Methodological critiques also show that online evaluations are not equivalent to traditional in person methods, challenging the assumption that data collected across modalities can be treated the same (Young & McCaslin, 2013; Zumrawi & Macfadyen, 2023). Administrators, faculty, and students are directly involved in this issue, with administrators relying on SET data for decision making, faculty affected by the outcomes of evaluations, and students shaping results through their responses. This paper will examine the causes of the problem by reviewing the recent literature on differences in SET outcomes across modalities, evaluating sources of bias, and analyzing the reliability and validity of SETs as a measure of teaching effectiveness.

Background of the Problem

The reliance on SETs has grown significantly over the past two decades as universities sought cost-effective and scalable ways to evaluate teaching quality (Benton & Cashin, 2012; Heffernan, 2023a). However, multiple studies have shown that SET results are inconsistent when comparing online and face-to-face courses. Research suggests that online courses often receive

lower ratings due to factors unrelated to teaching quality, such as technological issues, reduced personal interaction, and student preferences for in person contact (Lakeman et al., 2023; Wang et al., 2023). At the same time, face-to-face courses may benefit from immediacy and rapport effects that online courses cannot replicate (Daumiller et al., 2023). These discrepancies call into question the fairness of treating SET data as equivalent across modalities.

Moreover, the harmful outcomes of current practices extend beyond technical measurement issues. Faculty frequently report that SETs incentivize lenient grading and surface level teaching adjustments designed to "keep students happy," undermining rigor and academic standards (Lakeman, Coutts, Hutchinson, Massey, Nasrawi, Fielden, et al., 2022; Lakeman et al., 2023). SETs have also been shown to affect faculty mental health, particularly when negative or abusive comments are included in anonymous feedback (Heffernan, 2023a). Bias further compounds the problem, gender, race, and accent have all been found to influence SET results, meaning that marginalized academic often receive systematically lower scores regardless of teaching quality (Daskalopoulou, 2024; Zheng et al., 2023). When these biases intersect with instructional modality, for example, when women of international faculty teaching online, the disadvantages may be amplified.

Recent systematic reviews underscore the broader context: SET validity and reliability remain questionable, with sources of measurement errors and inconsistent rating behaviors identified across global studies (Quansah et al., 2024). Cross-national analyses also demonstrate that while SETs are used worldwide, their design and application differ across higher education traditions, challenging the assumption that one system can reliably capture teaching quality in all contexts (Buchanan et al., 2025; Spooren et al., 2013). Despite these concerns, universities continue to treat SETs as comparable across instructional formats, raising ethical and

administrative questions about equity and accuracy in faculty assessment. This study argues that higher education administrators must reconsider the use of SETs as a one-size-fits-all metric and instead evaluate teaching effectiveness in ways that account for modality differences and bias.

Thesis

This paper argues that student evaluations of teaching (SET), when applied equivalently across instructional modalities such as online and face-to-face courses, are fundamentally flawed due to issues of validity, systemic bias, modality-specific inequities, and harmful consequences for faculty and students; therefore, higher education institutions must adopt multi-measure, equitable, and transparent evaluation frameworks to ensure fairness, academic integrity, and alignment with their mission.

Literature Review

A substantial body of research has examined the use of student evaluations of teaching (SET) as measures of instructional effectiveness in higher education. While SETs remain a central tool in faculty assessment, the literature consistently reveals persistent problems that challenge their validity, fairness, and utility. Scholars have investigated how reliability and validity concerns undermine the accuracy of SETs, how systemic bias shapes outcomes in ways that disadvantage women and minority faculty, how instructional modality influences ratings in online versus face-to-face settings, and what consequences arise for faculty, students, and institutions when flawed measures are used in high-stakes decisions. This review synthesizes findings from recent scholarship, particularly within the past five years, while drawing on earlier studies to contextualize ongoing debates. The discussion is organized around four major themes, validity and reliability, bias, instructional modality, and consequences, that collectively illuminate the limitations of SETs and underscore the need for critical re-examination of their role in higher education.

Validity and Reliability of Student Evaluations of Teaching (SET)

Student evaluations of teaching (SET) have long been positioned as a central tool for assessing instructional effectiveness in higher education. Historically, SETs have been justified as cost-effective instruments capable of capturing students' perspectives on teaching quality (Benton & Cashin, 2012). They have been used to inform decisions on promotion, tenure, and contract renewal, and in many institutions, they remain one of the few standardized measures of teaching performance. However, the assumption that SETs are valid and reliable indicators of teaching quality has come under increasing scrutiny, particularly in the last decade.

Recent research highlights significant concerns about the reliability of SETs. Buchanan et al. (2025) found that reliability fluctuates depending on course size, grading practices, and contextual factors, raising doubts about whether SET scores consistently reflect teaching effectiveness. Their analysis, emphasizes that SETs are vulnerable to contextual noise, such as class size and student expectations, that distorts results (Buchanan et al., 2025). Similarly, Quansah et al. (2024), in a systematic review of global studies, concluded that the credibility of SET outcomes is questionable, largely due to measurement error and inconsistent student rating behaviors. They noted that students often fail to differentiate between distinct dimensions of teaching quality, instead providing global ratings influenced by unrelated factors such as grading leniency or course difficulty.

The question of validity, whether SETs actually measure what they purport to measure, has also been challenged. Daumiller et al. (2023) examined alignment between teacher self-reports and student evaluations across nearly 16,000 student assessments in Germany. They found only moderate agreement, particularly for less observable aspects of teaching such as cognitive activation and emotional support. This suggests that while SETs may capture certain surface-level elements (e.g., organization, clarity), they are weaker at detecting deeper

pedagogical qualities that contribute to learning outcomes. Their findings echo earlier critiques that SETs reflect student satisfaction rather than actual learning (Hornstein, 2017).

Psychometric perspectives reinforce these concerns. Sanchez et al. (2021) argue that many SET instruments lack construct validity due to insufficient psychometric testing. Using item response theory and multidimensional scaling, they demonstrate that current tools often collapse multiple dimensions of teaching into a single score, undermining their interpretive power. Zumrawi and Macfadyen (2023) further highlight that institutions rarely consider statistical best practices when summarizing SET data. They propose alternatives such as interpolated medians and dispersion indices to reduce the influence of outliers and skewed distributions, yet these methods remain underutilized in higher education practice.

Although new measurement models such as Generalizability Theory (GT) and Many-Facet Rasch Modeling (MFRM) have been applied to SET research, their use remains limited. Quansah et al. (2024) report that most recent studies in Asia and Africa still rely on classical test theory, which cannot adequately account for the multiple sources of error present in student ratings. Where GT and MFRM have been used, results show that the largest source of variability stems from the student rater rather than from the instructor or course, raising serious questions about validity (Quansah et al., 2024). This evidence suggests that SETs may be less a measure of teaching quality and more a reflection of student attitudes, predispositions, or situational factors.

SETs are neither fully reliable nor valid as measures of teaching quality. While they continue to be used widely across higher education, evidence increasingly points to the need for supplementary or alternative measures. The persistence of methodological weaknesses and measurement errors means that SETs cannot be considered a neutral or objective tool, particularly when used to make high-stakes decisions about faculty performance.

Bias in Student Evaluations

While questions of validity and reliability focus on whether student evaluations of teaching (SET) measure what they claim to measure, the issue of bias centers on whether certain groups of faculty are systematically disadvantaged by the process. A growing body of literature demonstrates that SETs are not neutral instruments. Instead, they reflect social and cultural prejudices that disproportionately affect women, racial and ethnic minorities, younger or older instructors, and faculty with non-native accents. These biases call into question the fairness of relying on SETs for high-stakes personnel decisions.

One of the most consistently documented sources of bias is gender. Zheng et al. (2023), analyzing over nine million reviews from RateMyProfessors, found significant disparities in the ways students evaluate male and female faculty in online environments. Female instructors were rated lower on dimensions such as authority and competence, while male instructors benefited from implicit credibility. Similarly, Mengel et al. (2019) found that male students rated female faculty up to 21% lower than male colleagues teaching the same courses. These findings align with MacNell et al. (2015) earlier experimental evidence showing that identical courses were rated more positively when students believed the instructor was male. Collectively, this research demonstrates that gender bias persists across modalities, with evidence suggesting it may be magnified in online contexts where personal interaction is limited.

Beyond gender alone, intersectional analyses reveal that multiple identity markers interact to shape SET outcomes. Daskalopoulou (2024), in a qualitative study of academics in the United Kingdom, found that faculty who identified as women, racial minorities, or disabled reported experiencing derogatory and abusive comments that targeted their identity rather than their teaching. Participants described the evaluations as detrimental to both mental health and career progression, with institutional structures often amplifying rather than mitigating these

harms. Fan et al. (2019) provide quantitative evidence that students give lower ratings to faculty from non-English-speaking backgrounds, particularly women of color, reinforcing the role of accent and cultural identity in bias. These studies show that marginalized academics experience "stacked disadvantages" in evaluation processes.

Other forms of bias have also been documented. Boring (2017) found that younger female instructors were often penalized for not meeting students' gendered expectations of warmth and nurturance, while older women reported being judged more harshly on appearance. More recent work by Heffernan (2023a) highlights the prevalence of abusive comments, particularly directed at women and marginalized groups, in open-ended SET responses. Such comments frequently focused on personal characteristics like body size, clothing, or accent rather than instructional effectiveness. The persistence of such non-constructive feedback undermines claims that SETs are objective measures of teaching.

Bias in SETs does not operate in isolation but has significant professional consequences. Studies consistently show that biased evaluations influence faculty promotion, tenure, and workload distribution. Lakeman, Coutts, Hutchinson, Massey, Nasrawi, Fielden, et al. (2022); Lakeman et al. (2023) document how the fear of negative evaluations leads faculty to adopt teaching strategies aimed at pleasing students rather than promoting rigorous learning. This performative approach not only diminishes educational quality but also disproportionately pressures those most vulnerable to bias. Daskalopoulou (2024) further found that women and minority academics were more likely to face stalled promotions and fewer leadership opportunities due to consistently lower evaluation scores. Thus, systemic bias embedded in SET processes becomes institutionalized in career advancement structures.

Another dimension of bias relates to the tone and content of narrative feedback. Cunningham et al. (2023) emphasize that abusive comments in SETs remain under-researched despite their clear harm. Their study advocates for automated systems to detect and filter abusive content, noting that institutions have largely failed to implement safeguards for faculty. Heffernan (2023a) similarly argues that universities underestimate the severity of verbal abuse in evaluations, particularly toward women and marginalized groups. Together, these findings suggest that SETs not only reproduce bias but also provide an institutional platform for harassment.

The literature converges on the conclusion that SETs are deeply affected by bias related to gender, race, intersectionality, age, and appearance. These biases undermine fairness, contribute to inequitable career outcomes, and distort the measurement of teaching effectiveness. Importantly, bias does not simply coexist with validity and reliability concerns; it compounds them, particularly in online modalities where anonymity and lack of personal contact may exacerbate prejudicial responses (Daskalopoulou, 2024; Zheng et al., 2023). Recognizing the systematic nature of these biases is essential for understanding why SETs, as currently implemented, fail to provide equitable or accurate assessments of teaching quality.

Instructional Modality and Student Evaluations of Teaching (SET)

One of the most critical but underexamined issues in student evaluations of teaching (SET) is whether results can be compared across instructional modalities. Many higher education institutions apply the same SET instruments to both online and face-to-face courses, assuming that teaching quality is measured equivalently in both settings. However, recent research demonstrates that modality influences how students perceive and rate instruction, which undermines the fairness and validity of treating scores as interchangeable.

A number of studies have compared SET outcomes across modalities, revealing systematic differences. Quansah et al. (2024) in their global systematic review, highlight that online courses often receive lower average ratings than face-to-face courses, even when course content and instructors remain the same. The lower scores, they argue, are linked less to instructional quality than to technological barriers, limited immediacy, and reduced interpersonal rapport. Daumiller et al. (2023) confirm that rapport and classroom climate strongly influence SET outcomes, and these factors are inherently more difficult to cultivate in online environments.

Earlier work provides additional context. Young and McCaslin (2013) questioned whether paper-based and online evaluation methods produce equivalent outcomes, noting that online systems tend to yield lower response rates and more polarized feedback. More recent evidence, however, suggests that differences are not merely methodological but stem from the modality of instruction itself. Wang et al. (2023) examining student perceptions of online evaluation systems, found that students are more critical when reviewing online courses, partly because the evaluation process emphasizes convenience rather than engagement.

Hybrid or blended courses introduce another layer of complexity. Lakeman et al. (2023) observe that instructors teaching in blended modalities often report evaluations shaped as much by technology integration as by pedagogy. Students penalize faculty for technological difficulties, even when these issues fall outside the instructor's control. This suggests that SETs in hybrid contexts may measure perceptions of institutional technology infrastructure as much as they measure instructional quality. Similarly, the study by Daumiller et al. (2023) found that "teaching presence" is rated more negatively in technology-mediated settings, especially when students feel disconnected from instructors.

Modality does not only shape ratings directly; it also interacts with known biases in student evaluations. Zheng et al. (2023) demonstrate that gender disparities in SETs are particularly pronounced in online settings. Women instructors teaching online received significantly harsher evaluations compared to men, with students more likely to question their competence in the absence of face-to-face interaction. Daskalopoulou (2024) further argues that marginalized academics may experience a compounding effect when teaching online, as anonymity in evaluations emboldens students to provide discriminatory or abusive comments. These findings highlight that instructional modality cannot be separated from broader concerns about equity in evaluation.

The evidence suggests that instructional modality is a significant factor influencing SET outcomes, yet institutions often ignore these differences in practice. When administrators treat online and face-to-face SET results as interchangeable, they risk making unfair judgments about faculty performance. As Zumrawi and Macfadyen (2023) caution, institutions often overlook methodological nuances in summarizing SET data, further compounding inequities across modalities. This practice undermines both the reliability of evaluation data and the fairness of personnel decisions that rely on them.

Research over the past five years indicates that SETs are strongly shaped by instructional modality. Online courses frequently receive lower evaluations due to factors unrelated to teaching quality, such as reduced rapport and technological challenges, while hybrid settings add further complications. Moreover, modality interacts with existing biases, often intensifying gender and racial disparities. These findings underscore the flawed assumption that SETs measure teaching equivalently across modalities and call for administrators to develop distinct, context-sensitive evaluation frameworks for online, hybrid, and face-to-face teaching.

Consequences of SET Use in Higher Education

The widespread reliance on student evaluations of teaching (SET) extends beyond questions of validity, reliability, bias, and modality. Scholars increasingly emphasize that the consequences of SET use for faculty, students, and institutions are profound. These consequences include impacts on faculty mental health and well-being, the shaping of instructional practices, the promotion of grade inflation, and the institutionalization of inequities in academic careers. Understanding these outcomes is critical for evaluating whether SETs serve the purposes for which they are intended.

Recent research highlights the negative psychological toll of SETs on instructors, particularly when evaluations contain biased or abusive comments. Heffernan (2023) documents how women and marginalized faculty are disproportionately subjected to derogatory remarks, which not only undermine professional credibility but also contribute to stress and emotional exhaustion. Similarly, Lakeman, Wright-Brough, and Sargent (2022) found that anonymous SET systems contribute to high levels of distress, anxiety, and depressive symptoms among faculty in Australian universities. Their participants reported that repeated exposure to negative or abusive feedback fostered a sense of vulnerability, diminished confidence, and, in some cases, clinical mental health issues. These findings highlight that SETs are not neutral instruments; they can function as vehicles of harm, particularly in environments where faculty lack institutional protections.

A recurring theme in the literature is that SETs create incentives for faculty to prioritize student satisfaction over learning outcomes. Stark and Freishtat (2014) argue that evaluations often correlate more strongly with expected or received grades than with actual learning. More recent evidence by Lakeman et al. (2023) confirms that instructors may adjust course content, reduce assessment difficulty, or avoid pedagogical innovation to avoid negative evaluations.

Johnson (2003) previously observed that lenient grading often yields higher evaluations, a dynamic that persists today. The cumulative effect is a form of "teaching to the evaluation," where maintaining student approval takes precedence over fostering deep learning. This consequence raises concerns about academic integrity and the long-term educational mission of higher education institutions.

Because SETs are often central to faculty evaluation processes, biases and flawed measurements translate directly into career inequities. Daskalopoulou (2024) shows that biased evaluations negatively impact promotion, tenure, and leadership opportunities, particularly for women and minority academics. Uttl and Simbert (2017) add that relying on SETs for personnel decisions may explain, in part, the persistent underrepresentation of women in senior academic positions. When evaluation scores are used as benchmarks for probation, renewal, or promotion, marginalized faculty disproportionately face stalled career advancement, perpetuating systemic inequities. This institutionalization of bias is particularly problematic given that many universities frame SETs as evidence of teaching quality in promotion dossiers and merit reviews.

The consequences of SET use extend to students as well. Marshik et al. (2023) argue that when institutions overly emphasize student satisfaction, teaching can devolve into performance designed to entertain rather than educate. Students may come to view themselves as customers whose preferences must be satisfied, rather than as learners engaged in a rigorous academic process. Wang et al. (2023) further note that online SET systems often reinforce this consumerist model by focusing on convenience and usability, rather than fostering meaningful dialogue about learning. At the policy level, Spooren et al. (2013) observe that universities often adopt SETs for accountability and benchmarking purposes, even when evidence of validity is weak. This

institutional emphasis reflects broader neoliberal trends in higher education, where market-driven models prioritize efficiency and customer satisfaction over educational quality.

Globally, the consequences of SET use vary, but certain patterns recur. Quansah et al. (2024) highlight that in regions such as Africa and Asia, where student feedback is increasingly integrated into quality assurance frameworks, the same concerns about bias and validity apply. Buchanan et al. (2025) emphasize that reliability challenges persist across contexts, suggesting that the limitations of SETs are not culture-specific but inherent to the tool itself. At the same time, The Debate on Student Evaluations of Teaching (Spooren et al., 2013) notes that the ways SETs are framed, whether as accountability measures, professional development tools, or instruments for student voice, differ internationally, shaping the consequences for faculty careers and institutional governance. These variations underscore that while SET use is global, its effects are conditioned by local higher education traditions and policies.

The literature indicates that the consequences of SET use are far-reaching and often negative. Faculty face psychological harm, distorted incentives, and career inequities, while students may experience a commodified form of education that privileges satisfaction over learning. Institutions, in turn, risk perpetuating systemic inequities and undermining academic standards by over-relying on flawed instruments. Taken together, these findings strengthen the case that SETs, particularly when used equivalently across instructional modalities, are ill-suited to serve as high-stakes measures of teaching quality in higher education. Addressing these consequences requires not only methodological reform but also a rethinking of the broader role that student evaluations should play in academic governance.

Analysis

The literature review identified four recurring themes: (1) validity and reliability concerns, (2) systemic bias, (3) instructional modality differences, and (4) consequences for stakeholders. To analyze these themes, this section applies four theoretical frameworks, Institutional Theory, Critical Theory, Role Theory, and Expectancy Theory, while considering the issue from three perspectives: organizational/administrative, faculty/individual, and student/learner. Together, these lenses reveal the systemic, structural, and motivational dynamics that shape the use and impact of student evaluations of teaching (SET).

Institutional Theory

Institutional Theory provides a powerful framework for understanding why student evaluations of teaching (SET) continue to dominate higher education assessment practices despite decades of evidence highlighting their flaws. Institutional Theory explains how organizations adopt and maintain practices not necessarily because they are the most effective, but because they confer legitimacy, stability, and alignment with broader societal norms (DiMaggio & Powell, 1983; Meyer & Rowan, 1977). Central to this perspective is the concept of institutional isomorphism, which suggests that organizations in the same field tend to resemble one another over time as they respond to pressures for conformity, legitimacy, and efficiency. Practices become taken-for-granted, even when evidence questions their effectiveness, because they serve symbolic functions that signal accountability to external stakeholders.

Applied to higher education, Institutional Theory helps explain the persistence of student evaluations of teaching (SET). Despite decades of research highlighting flaws in validity and reliability, SETs remain widespread because they are viewed as standardized, cost-effective, and comparable measures of teaching quality (Buchanan et al., 2025; Spooren et al., 2013). Universities adopt and retain SETs less because they are empirically sound and more because

they are perceived as legitimate markers of accountability and responsiveness to students (Quansah et al., 2024). Once SETs became a normative practice in the United States during the mid-20th century, they quickly spread internationally as a symbol of accountability and responsiveness to students Spooren et al. (2013). Today, SETs are nearly universal across higher education systems worldwide, from North America and Europe to Asia and Africa (Quansah et al., 2024). Institutional Theory helps explain this persistence: rather than being based on strong evidence of validity, the continued use of SETs is driven by conformity to external expectations, accreditation standards, and perceived legitimacy.

Validity and Reliability

From an administrative perspective, validity and reliability concerns are often downplayed because SETs provide a steady stream of standardized data that can be easily aggregated and reported. As Buchanan et al. (2025) point out, administrators often treat SET scores as objective benchmarks despite their vulnerability to measurement error. Institutions adopt SETs not because they are flawless tools but because they satisfy demands for efficiency and comparability.

From a faculty perspective, this institutionalization means that professors are evaluated by tools that do not necessarily reflect their teaching quality. Quansah et al. (2024) show that SETs suffer from inconsistent reliability across cultural contexts, yet faculty careers are nonetheless judged by their results. Faculty frequently express frustration that institutional leaders ignore the methodological critiques well-documented in scholarly research, reinforcing a sense of powerlessness within the system (Daskalopoulou, 2024).

From a student perspective, the institutionalization of SETs frames evaluations as the primary channel for providing input into teaching quality. Yet the standardized nature of SET instruments often strips away nuance, leaving students with generic rating scales that may not

capture the complexity of their experiences (Wang et al., 2023). This contributes to a perception of SETs as a procedural formality rather than a meaningful feedback mechanism.

Bias

Institutional Theory also helps explain how bias becomes embedded in faculty evaluation systems. From the administrative perspective, bias is often ignored because SET scores provide convenient metrics for decision-making. By treating all evaluations as equal, institutions inadvertently legitimize biased results (Zheng et al., 2023).

From a faculty perspective, this institutional blindness translates into systemic disadvantage. Women, racial minorities, and non-native English speakers face persistent bias, yet institutions rarely adjust SET results to account for these disparities (Daskalopoulou, 2024). Faculty members see this as evidence that institutional priorities, efficiency, comparability, and reputation, take precedence over equity and fairness.

From a student perspective, institutional practices shape how feedback is used. Students may believe their evaluations are purely about teaching quality, yet the way institutions aggregate and apply results reinforces patterns of inequity. Critical findings, such as those by Heffernan (2023a) on abusive comments, are often downplayed in institutional reports, allowing bias to persist unchecked.

Modality

Instructional modality is another area where Institutional Theory sheds light on organizational inertia. From an administrative perspective, treating online and face-to-face evaluations as equivalent allows for comparability across programs and modalities. Institutions often resist creating separate frameworks for online teaching, citing concerns about cost and complexity (Zumrawi & Macfadyen, 2023).

From a faculty perspective, this oversight is highly consequential. Online instructors consistently receive lower ratings due to reduced immediacy and technological frustrations, yet institutions fail to differentiate between modality effects and actual teaching quality (Daumiller et al., 2023). Faculty perceive this as an institutional failure to recognize the distinctive challenges of online instruction.

From a student perspective, modality differences are often invisible in the reporting of results. Students may assume their feedback is treated fairly, but institutions' failure to account for modality creates systemic distortions in how results are interpreted and applied.

Consequences

Finally, the consequences of SET use can be analyzed through the lens of Institutional Theory. For administrators, SETs provide a veneer of accountability that satisfies external stakeholders such as accreditation agencies, policymakers, and boards of trustees. For faculty, however, the consequences include stalled promotions, mental health strain, and distorted teaching practices (Heffernan, 2023a; Lakeman, Coutts, Hutchinson, Massey, Nasrawi, Fielden, et al., 2022)For students, the institutionalization of SETs fosters a consumerist model of education, reinforcing the idea that satisfaction is equivalent to learning outcomes (Marshik et al., 2023).

Institutional Theory thus reveals how organizational inertia and legitimacy-seeking behaviors drive the persistence of SETs despite validity concerns, bias, modality effects, and harmful consequences. Institutions perpetuate flawed practices because they serve symbolic and political functions, even when evidence shows they undermine equity and teaching quality.

Critical Theory

Critical Theory, originating with the Frankfurt School, is concerned with exposing and challenging systems of domination, inequality, and distorted power relations in society

(Habermas, 1984; Horkheimer, 1972). It argues that institutional practices often appear neutral but, in reality, reproduce existing hierarchies and social injustices. Applied to education, Critical Theory highlights how structures such as assessment systems can privilege some groups while marginalizing others, thereby perpetuating inequities under the guise of objectivity (Brookfield, 2005; Kincheloe & McLaren, 2002). In the context of student evaluations of teaching (SET), Critical Theory provides a lens to examine how validity concerns, bias, modality differences, and systemic consequences reflect not only methodological weaknesses but also deeper patterns of structural inequality embedded within higher education.

Validity and Reliability

When viewed through the lens of validity and reliability, Critical Theory demonstrates how institutions privilege efficiency and accountability over fairness and evidence. From an administrative perspective, validity concerns are often dismissed because SETs produce standardized quantitative data that align with neoliberal models of accountability (Quansah et al., 2024; Spooren et al., 2013). Institutions frame SETs as objective even when research consistently shows they are unreliable and methodologically weak. This framing maintains managerial legitimacy at the expense of fairness. For faculty, the continued reliance on invalid scores in promotion and tenure decisions is experienced as systemic injustice. As Buchanan et al. (2025) note, SET scores remain unreliable across contexts, yet faculty careers are nonetheless shaped by them. This subordinates academic expertise to managerial priorities, eroding trust between instructors and their institutions. From the student perspective, validity flaws signal that their feedback is commodified into numerical ratings that serve administrative reporting needs rather than meaningful dialogue. Students' voices are simplified into data points, which reinforces institutional authority rather than facilitating authentic educational improvement (Marshik et al., 2023).

Bias

Bias is perhaps the most clearly illuminated theme under Critical Theory, exposing how discrimination is embedded in institutional practices. Administratively, biased outcomes are often ignored or minimized, with institutions treating results as neutral despite abundant evidence of gender and racial disparities (Zheng et al., 2023). This embeds inequity directly into decision-making systems. For faculty, especially women, racial minorities, and non-native English speakers, bias is experienced as abusive comments, lower scores, and stalled career progression (Heffernan, 2023a). Critical Theory interprets these not as incidental flaws but as structural injustices, where prejudice is legitimized through official evaluation systems. From the student perspective, evaluations function as vehicles for reproducing societal stereotypes. Studies reveal that students judge women more harshly and penalize faculty who do not conform to gendered expectations of authority or warmth (Boring, 2017; MacNell et al., 2015). In this way, student evaluations reflect and reproduce inequality when incorporated into institutional processes.

Modality

Modality further illustrates how SETs perpetuate inequity. Administratively, treating online and face-to-face evaluations as equivalent erases the distinct disadvantages faced by online instructors. Institutions prefer comparability and cost-efficiency over fairness, thereby institutionalizing inequities across modalities (Zumrawi & Macfadyen, 2023). Faculty experience this erasure most acutely. Research shows that gender bias intensifies in online evaluations, and anonymity in digital spaces emboldens students to leave discriminatory or abusive comments (Daskalopoulou, 2024; Zheng et al., 2023). Rather than providing a fairer evaluation environment, online platforms replicate and even exacerbate the inequalities of traditional settings. Students, meanwhile, may perceive online evaluations as feedback

opportunities, but Critical Theory demonstrates that anonymity fosters discriminatory expression and harsher judgments. In this way, instructional modality becomes an amplifier of inequity.

Consequences

The consequences of SET use are also illuminated through a critical lens. For administrators, the continued reliance on SETs reflects neoliberal logics that frame students as customers and emphasize satisfaction over learning outcomes (Fleming, 2019). This redefines teaching as a transactional service rather than an intellectual endeavor, aligning institutional practices with consumerist expectations. For faculty, the consequences are substantial: biased evaluations lead to stalled promotions, job insecurity, stress, and psychological harm (Heffernan, 2023a; Lakeman, Coutts, Hutchinson, Massey, Nasrawi, Fielden, et al., 2022). Marginalized groups are disproportionately affected, reinforcing systemic underrepresentation of women and minorities in senior academic positions (Uttl & Simbert, 2017). Students also face consequences, though in subtler ways. When SETs reward leniency and entertainment rather than rigor, students receive an education shaped by consumer logic rather than intellectual challenge (Stark & Freishtat, 2014). This undermines higher education's transformative potential and reproduces inequities in access to meaningful learning.

Critical Theory reveals that SETs are more than flawed measurement tools; they are mechanisms that reinforce structural inequities and perpetuate power imbalances. By embedding bias, disregarding validity concerns, ignoring modality effects, and producing harmful consequences, institutions reproduce injustices under the guise of accountability and student voice. For administrators, SETs legitimize managerial control; for faculty, they function as systemic oppression; and for students, they reflect and amplify social prejudices. Unless institutions radically reform these practices, SETs will continue to perpetuate inequity and compromise the mission of higher education.

Role Theory

Role Theory examines how individuals understand, perform, and negotiate the social roles they occupy, along with the expectations attached to those roles (Biddle, 1986; Turner, 2001). It posits that much of human behavior is shaped by socially defined categories, such as teacher, student, or administrator, and the norms, obligations, and expectations associated with those roles. Role conflict arises when the demands of different roles are inconsistent or contradictory, while role strain occurs when an individual struggles to meet expectations within a single role (Goode, 1960). Within higher education, Role Theory provides an effective framework for understanding how student evaluations of teaching (SET) intensify tensions between institutional expectations, faculty identity, and student perceptions.

Applied to SETs, Role Theory highlights how administrators, faculty, and students assign and interpret roles in ways that contribute to validity concerns, reproduce bias, complicate modality, and shape broader consequences. Faculty members, in particular, experience SETs as a source of role conflict: they are expected to be rigorous educators and simultaneously approachable service providers. Students often evaluate based on the latter expectations, while administrators frequently reward conformity to satisfaction metrics rather than pedagogical integrity.

Validity and Reliability

Through the lens of Role Theory, validity and reliability issues in SETs can be understood as stemming from misaligned role expectations. Administrators often assume that faculty roles can be accurately assessed by students through standardized surveys. Yet research demonstrates that student judgments are influenced more by satisfaction with service-oriented aspects of the role, such as responsiveness or leniency, than by teaching effectiveness (Buchanan et al., 2025; Stark & Freishtat, 2014). Faculty, therefore, encounter role strain, as they are held

accountable to measures that privilege aspects of their role peripheral to actual pedagogy. From the student perspective, the role of "evaluator" is shaped less by academic expertise and more by consumer expectations. As Quansah et al. (2024) note, this weakens reliability, as students apply different criteria based on personal experiences rather than consistent standards of teaching quality.

Bias

Bias within SETs is also illuminated by Role Theory, which shows how gender, race, and cultural stereotypes shape expectations of faculty roles. Administratively, institutions often interpret evaluations as neutral reflections of performance, but in reality, they reinforce stereotypical role expectations—for instance, that women instructors should be nurturing or that male instructors should be authoritative (Boring, 2017; MacNell et al., 2015). Faculty experience this bias as role conflict: women may be penalized for being too assertive while men are rewarded for the same behavior (Zheng et al., 2023). Marginalized faculty must negotiate conflicting role expectations that systematically disadvantage them (Daskalopoulou, 2024). From the student perspective, evaluations often reflect whether instructors met socially constructed role expectations rather than whether they were effective educators. This explains why bias persists so strongly in SET results: it is rooted in the misalignment between socially prescribed faculty roles and professional teaching roles.

Modality

Instructional modality compounds these role dynamics. Administrators, seeking comparability, often treat online and face-to-face SETs as equivalent, overlooking how modality alters the faculty role. Online instructors, for example, cannot perform traditional relational cues of warmth and authority in the same ways as in-person instructors. Faculty teaching online therefore experience intensified role strain, as students expect immediacy and rapport that the

medium constrains (Daumiller et al., 2023). For marginalized faculty, these strains intersect with bias, since anonymity in online settings emboldens discriminatory evaluations (Zheng et al., 2023). From the student perspective, online courses shift their expectations of faculty roles: they may view instructors more as service providers responsible for technological smoothness than as educators. As Zumrawi and Macfadyen (2023) caution, institutions' failure to account for modality differences exacerbates role conflict and undermines fair evaluation.

Consequences

The consequences of SETs can also be understood as consequences of role misalignment. Administrators, driven by accountability pressures, often reward faculty who fulfill service-oriented role expectations (e.g., leniency, availability) rather than those who challenge students with rigor (Fleming, 2019). Faculty, in turn, may engage in grade inflation or adjust teaching strategies to align with student expectations, even when doing so undermines pedagogical goals (Lakeman et al., 2023). This reflects role conflict between being rigorous educators and satisfying evaluators. Students, meanwhile, experience education increasingly framed as a consumer transaction, reinforcing their role as customers rather than learners (Marshik et al., 2023). This shifts the meaning of education away from intellectual growth and toward service satisfaction.

Role Theory demonstrates that SETs perpetuate tensions among administrators, faculty, and students by embedding contradictory role expectations into evaluation systems. For administrators, SETs reinforce a managerial role that prizes efficiency and satisfaction. For faculty, they intensify role conflict and strain, forcing instructors to navigate between rigor and popularity. For students, evaluations shape and are shaped by consumerist role expectations, positioning them as customers rather than co-learners. By examining validity issues, bias, modality differences, and consequences through this lens, it becomes clear that SETs do not

simply measure teaching quality but instead reproduce and exacerbate role misalignments across higher education.

Expectancy Theory

Expectancy Theory, developed by Victor Vroom (1964)is a motivational framework that explains how individuals make decisions based on the expected outcomes of their actions. The theory posits that motivation is determined by three factors: expectancy (belief that effort leads to performance), instrumentality (belief that performance leads to outcomes), and valence (the value placed on those outcomes). In educational contexts, Expectancy Theory has been used to explain student engagement, decision-making, and evaluation behaviors (Van Eerde & Thierry, 1996). Students' evaluations of faculty are shaped by their expectations of what teaching should deliver, including grades, workload, and satisfaction. When these expectations are not met, evaluations often reflect disappointment rather than teaching effectiveness.

Applied to student evaluations of teaching (SET), Expectancy Theory provides a valuable lens for understanding how students' motivations distort evaluation results, how faculty adjust teaching practices to align with expected outcomes, and how administrators interpret evaluations as signals of institutional performance. This framework helps explain why validity problems, bias, modality effects, and systemic consequences persist in the use of SETs.

Validity and Reliability

Expectancy Theory highlights that the validity of SETs is undermined by the misalignment between what students expect and what institutions intend to measure. Administrators often assume that SETs reflect instructional quality, but students frequently base evaluations on whether instructors met their expectations for leniency, workload, or grades (Stark & Freishtat, 2014). Faculty recognize that these expectations shape student judgments, creating incentives to adjust teaching practices to align with what will produce favorable evaluations,

even if such adjustments undermine rigor (Lakeman et al., 2023). For students, evaluations become a reflection of whether their personal valence, the value they place on receiving higher grades or lighter workloads, was fulfilled. This explains why SET scores often correlate more strongly with anticipated or actual grades than with learning outcomes (Uttl & Simbert, 2017), raising serious questions about validity and reliability.

Bias

Expectancy Theory also sheds light on how bias emerges in SETs. Administrators may interpret low scores as signs of poor teaching, without considering that student expectations vary based on stereotypes about gender, race, or accent. Faculty from marginalized groups face a double bind: students often expect women to be more nurturing or minorities to be more accommodating, and when those expectations are not met, evaluations suffer (Boring, 2017; Daskalopoulou, 2024). For faculty, this creates pressure to conform to biased expectations rather than pedagogical best practices. From the student perspective, biases operate through expectations of role performance: male instructors may be assumed competent until proven otherwise, while female or minority instructors must work harder to meet students' stereotyped benchmarks of authority or warmth (MacNell et al., 2015). Expectancy Theory thus illustrates how differential expectations translate into systematically biased outcomes.

Modality

Instructional modality further complicates expectations. Administrators tend to assume that online and face-to-face evaluations are equivalent, but Expectancy Theory reveals that students bring different expectations to each modality. In face-to-face courses, students often expect immediacy, rapport, and dynamic delivery, while in online courses, they expect technological fluency, rapid responses, and seamless learning platforms (Daumiller et al., 2023). When these modality-specific expectations are not met, evaluations suffer, regardless of

instructional quality. Faculty teaching online therefore experience strong pressure to meet technological and service expectations that may fall outside their teaching role, while students evaluate based on the gap between expected and actual experiences. Zheng et al. (2023) further demonstrate that in online settings, expectations are shaped by gender stereotypes, with women penalized more harshly when they fail to meet perceived standards of responsiveness or support.

Consequences

The consequences of SET use, when viewed through Expectancy Theory, show how misaligned motivations distort higher education. Administrators rely on evaluations as signals of teaching quality, yet they often measure satisfaction rather than learning. This leads institutions to reinforce policies that reward high evaluation scores without interrogating the expectations behind them (Fleming, 2019). For faculty, the consequences include strategic teaching behaviors, such as grade inflation, softened assessments, or increased availability, that align with student expectations but compromise academic rigor (Lakeman et al., 2023). For students, the consequence is an education increasingly shaped by consumerist expectations: evaluations reward instructors who meet demands for convenience and satisfaction, creating a feedback loop that devalues intellectual challenge (Marshik et al., 2023). Over time, this dynamic risks undermining the integrity of higher education, as teaching effectiveness becomes conflated with the fulfillment of expectations rather than meaningful learning outcomes.

Expectancy Theory explains why SETs so often reflect student satisfaction rather than instructional quality. Students evaluate based on whether their expectations were met, faculty adapt to secure positive outcomes, and administrators treat evaluations as institutional performance metrics. The result is a system where validity and reliability are compromised, biases are reinforced, modality differences are amplified, and consequences reshape both teaching and learning. From this perspective, SETs do not simply measure teaching

effectiveness, they measure the alignment between expectations and experiences, a dynamic that undermines their use in high-stakes decisions across higher education.

Ethical Implications

The continued reliance on student evaluations of teaching (SET) as high-stakes measures of faculty performance raises serious ethical concerns in higher education. If validity and reliability concerns, systemic bias, modality differences, and harmful consequences are left unaddressed, the ethical foundations of academic evaluation and decision-making are undermined. At stake are issues of fairness, equity, accountability, and the very integrity of the educational mission.

Fairness and Validity

At the core of ethical evaluation is fairness: decisions about faculty careers should be based on accurate and reliable information. Yet extensive research demonstrates that SETs lack validity and reliability, with results influenced more by student satisfaction, course difficulty, or anticipated grades than by teaching effectiveness (Buchanan et al., 2025; Quansah et al., 2024; Stark & Freishtat, 2014). Continuing to use SETs without addressing these shortcomings violates the principle of justice, as faculty may be rewarded or penalized based on data that do not accurately reflect their performance. For administrators, knowingly relying on flawed measures raises ethical questions about transparency and accountability. Students are also ethically implicated when their feedback, collected under the assumption that it measures teaching quality, is misused for purposes beyond their intent.

Equity and Bias

The ethical stakes are particularly stark in relation to bias. Evidence shows that women, racial and ethnic minorities, and non-native English speakers consistently receive lower

evaluations due to stereotypes and prejudice (Boring, 2017; MacNell et al., 2015; Zheng et al., 2023). The persistence of abusive comments directed at marginalized faculty further compounds this inequity (Heffernan, 2023a). When biased evaluations are treated as objective indicators of teaching quality, institutions perpetuate systemic discrimination under the guise of neutrality. From an ethical standpoint, this contravenes commitments to diversity, equity, and inclusion, effectively embedding injustice into faculty evaluation processes. For faculty, the consequences include stalled career progression, reduced well-being, and diminished representation in leadership roles (Daskalopoulou, 2024). For students, the ethical cost lies in being complicit in reinforcing stereotypes through institutional structures.

Modality and Justice

The ethical implications of modality differences also demand attention. Online courses often receive lower ratings than face-to-face courses, even when content and instructors are equivalent (Daumiller et al., 2023; Quansah et al., 2024). Treating these results as comparable disregards the distinct challenges of online instruction, thereby disadvantaging faculty who teach in digital environments. This is especially concerning as online education continues to expand globally. When institutions ignore modality differences, they risk creating a two-tiered system where faculty who teach online are systematically disadvantaged in evaluations, promotions, and job security. Such practices violate principles of equity and distributive justice, as faculty are penalized for factors beyond their control, including institutional technology infrastructure. For students, this inequity also has ethical implications: if online instructors are evaluated unfairly, institutions may deprioritize investments in online education, thereby disadvantaging students who rely on flexible or distance learning opportunities.

Consequences for Well-Being and Academic Integrity

The ethical consequences of SET use extend to both individual well-being and broader academic integrity. Faculty frequently report that negative evaluations cause stress, anxiety, and diminished confidence, with some experiencing clinical impacts on mental health (Lakeman, Coutts, Hutchinson, Massey, Nasrawi, & Fielden, 2022). Institutions that fail to protect faculty from abusive or biased evaluations abdicate their ethical responsibility to safeguard employee well-being. Beyond individual harm, SETs incentivize grade inflation and "teaching to the evaluation," undermining the rigor and integrity of higher education (Lakeman et al., 2023; Stark & Freishtat, 2014). Students, in turn, are ethically harmed when their education is shaped by consumer satisfaction rather than intellectual growth. Institutions that prioritize satisfaction scores over learning outcomes risk betraying their mission to foster critical thinking and prepare graduates for complex societal challenges.

Institutional Responsibility

Ultimately, the ethical implications of SETs highlight the responsibility of institutions to critically evaluate and reform their practices. Continuing to use SETs as the primary or sole measure of teaching effectiveness perpetuates injustice, undermines equity, and erodes trust between faculty, students, and administration. From an ethical standpoint, institutions must adopt more holistic and triangulated approaches to evaluation, such as peer observation, self-assessment, and learning outcome measures (Benton & Cashin, 2012; Berk, 2005). Failure to do so signals complicity in systemic inequity and a disregard for the principles of fairness, integrity, and inclusion that higher education claims to uphold.

Ethical Implications Summary

If unaddressed, the problems associated with SETs pose profound ethical risks. Faculty are judged by invalid and biased instruments, marginalized groups bear disproportionate harms, online instructors face structural disadvantages, and the overall quality of education is

compromised by distorted incentives. For administrators, the ethical challenge is one of accountability: to recognize the harm perpetuated by flawed practices and to reform evaluation systems accordingly. For faculty, the stakes involve not only career trajectories but also well-being and professional dignity. For students, the ethical implications include complicity in perpetuating inequities and receiving an education increasingly shaped by consumerist values. Addressing these ethical implications is not optional but necessary if higher education is to remain true to its mission of equity, integrity, and intellectual growth.

Policy Recommendations

The evidence reviewed demonstrates that student evaluations of teaching (SET) in their current form are unreliable, biased, and ethically problematic when used as high-stakes measures of faculty performance. To address these issues, higher education institutions must adopt comprehensive policy reforms that emphasize fairness, accuracy, and equity. The following recommendations highlight strategies that can mitigate the limitations of SETs while promoting more holistic approaches to evaluating teaching effectiveness.

Adopt a Multi-Measure Framework for Teaching Evaluation

The first and most widely endorsed policy recommendation is to move beyond sole reliance on SETs. Scholars emphasize the importance of triangulating multiple sources of evidence, including peer review of teaching, self-reflection, and measures of student learning outcomes (Benton & Cashin, 2012; Berk, 2005). Institutions should require that SETs comprise only a portion of a broader portfolio, ensuring that no single metric can disproportionately affect faculty careers. For example, Benton and Cashin (2012) argue that SETs are most useful when combined with peer observation and self-assessment. Marshik et al. (2023) similarly call for universities to design instruments that capture nuanced aspects of teaching quality rather than

focusing exclusively on student satisfaction. By embedding SETs in a multi-measure framework, institutions can improve validity and reduce the risk of unfair or distorted evaluations.

Address Bias Through Institutional Safeguards

To mitigate systemic inequities, institutions must implement policies that identify, monitor, and adjust for bias in SET data. Research consistently shows that women and minority faculty are disadvantaged by biased evaluations (Boring, 2017; Daskalopoulou, 2024; Zheng et al., 2023). One strategy is to include disclaimers in evaluation forms informing students of common biases and reminding them to focus on teaching practices rather than personal characteristics (Heffernan, 2023a). Automated systems can also detect and filter abusive comments, as suggested by Cunningham et al. (2023) ensuring that harmful feedback is not passed to faculty. At the administrative level, data analytics should be used to identify systematic disparities in scores across demographic groups and modalities. Institutions must then adjust policies to ensure that biased outcomes do not directly influence promotion or tenure decisions.

Differentiate Evaluation Practices by Instructional Modality

Given evidence that instructional modality shapes SET outcomes (Daumiller et al., 2023; Quansah et al., 2024), institutions should adopt distinct evaluation frameworks for online, hybrid, and face-to-face teaching. Online instructors should not be disadvantaged by factors such as technological infrastructure or reduced opportunities for immediacy, which are largely beyond their control. Zumrawi and Macfadyen (2023) argue for revised statistical metrics, such as interpolated medians, to account for skew and modality-related variance in SET data. Moreover, institutions should develop modality-specific instruments that assess unique dimensions of online and hybrid teaching, including clarity of communication, accessibility of materials, and

effectiveness of digital engagement tools. Differentiating evaluations by modality ensures more equitable treatment of faculty and produces data better aligned with instructional contexts.

Prioritize Faculty Well-Being and Academic Integrity

Institutions have an ethical responsibility to protect faculty well-being and ensure that evaluation practices support academic integrity. Research demonstrates that SETs often cause stress, anxiety, and harm to faculty mental health (Heffernan, 2023a; Lakeman, Coutts, Hutchinson, Massey, Nasrawi, Fielden, et al., 2022). To address this, institutions should establish policies that prohibit the use of abusive comments in formal evaluations and provide support resources for faculty experiencing evaluation-related distress. Additionally, to counter incentives for grade inflation and "teaching to the evaluation," institutions should explicitly emphasize rigor and student learning outcomes in faculty assessments (Lakeman et al., 2023; Stark & Freishtat, 2014). Encouraging peer mentoring, teaching development programs, and reflective practice can further shift evaluation systems toward continuous improvement rather than punitive measures.

Enhance Transparency and Accountability in Evaluation Use

Finally, institutions must be transparent about how SET data are collected, analyzed, and applied in decision-making. Faculty should have access to clear explanations of how their evaluations are interpreted, including whether adjustments are made for response rates, modality, or systemic bias. As Berk (2005) argues, transparency is critical for fostering trust in evaluation systems. Institutions should also engage students in the redesign of evaluation tools, helping them understand how their feedback is used and emphasizing the responsibility to provide constructive and equitable evaluations. By embedding accountability in the evaluation process, institutions can better align practices with their stated commitments to fairness, inclusion, and integrity.

Policy Recommendations Summary

Policy reforms must move beyond incremental adjustments to SET instruments and address the systemic problems embedded in their use. Adopting multi-measure frameworks, addressing bias through safeguards, differentiating by instructional modality, prioritizing faculty well-being, and enhancing transparency are essential steps toward equitable and valid evaluation systems. Implementing these recommendations requires institutions to confront entrenched practices and embrace a more holistic and ethical approach to teaching evaluation. Without such changes, SETs will continue to perpetuate inequities, distort incentives, and undermine the integrity of higher education.

Conclusion

Student evaluations of teaching (SET) remain one of the most widely used instruments for assessing teaching effectiveness in higher education. Despite their ubiquity, this paper has demonstrated that SETs are deeply flawed when applied equivalently across instructional modalities, such as online and face-to-face courses. While institutions continue to rely on these tools for promotion, tenure, and faculty development decisions, the evidence from contemporary research overwhelmingly shows that SETs lack validity, reproduce systemic bias, distort teaching practices, and negatively affect both faculty and students. By integrating findings from recent literature, analyzing themes through multiple theoretical frameworks, considering ethical implications, and outlining policy recommendations, this paper argues that without meaningful reform, SETs will continue to undermine the mission and values of higher education.

The central problem addressed in this paper is that student evaluations of teaching are treated as equivalent across different instructional modalities despite mounting evidence that this approach is flawed and unreliable. Online courses consistently receive lower ratings than face-to-

face courses due to factors such as reduced immediacy, technological barriers, and shifting student expectations (Daumiller et al., 2023; Quansah et al., 2024). Institutions that ignore these differences effectively disadvantage faculty who teach in online contexts. Furthermore, SETs are compromised by long-standing issues of validity and reliability, systemic bias based on gender and race, and consequences that extend far beyond measurement error. The persistence of these problems raises urgent questions about fairness, equity, and integrity in higher education.

The literature review revealed four major themes. First, concerns about validity and reliability undermine the use of SETs as objective measures. Research demonstrates that SET scores correlate more strongly with grades and student satisfaction than with actual learning outcomes, making them a poor proxy for teaching effectiveness (Buchanan et al., 2025; Stark & Freishtat, 2014). Second, bias consistently skews results against women, minority faculty, and non-native English speakers, with abusive and non-constructive comments compounding inequities (Heffernan, 2023a; Zheng et al., 2023). Third, instructional modality plays a significant role in shaping outcomes, with online courses penalized by factors largely outside instructors' control (Zumrawi & Macfadyen, 2023). Fourth, the consequences of reliance on SETs include harm to faculty well-being, incentives for grade inflation, erosion of academic rigor, and reinforcement of consumerist models of education (Lakeman, Coutts, Hutchinson, Massey, Nasrawi, Fielden, et al., 2022; Marshik et al., 2023). Taken together, these findings reveal that SETs, as currently implemented, fail to provide a fair, valid, or equitable measure of teaching.

Analyzing the problem through four theoretical frameworks provided deeper insight into why these flaws persist and how they operate. Institutional Theory explained why SETs remain entrenched despite evidence of their limitations. Once established as a norm, SETs became

institutionalized practices that signal legitimacy and accountability to external stakeholders, even if they lack effectiveness (DiMaggio & Powell, 1983; Meyer & Rowan, 1977). Critical Theory illuminated how SETs function as mechanisms of systemic inequity. By embedding gender and racial bias into official evaluation systems, SETs legitimize and perpetuate structural discrimination under the guise of neutrality (Brookfield, 2005; Daskalopoulou, 2024).

Role Theory shed light on the role conflicts faculty experience when confronted with competing expectations. Instructors are expected to be rigorous educators and simultaneously service providers who cater to student satisfaction, a contradiction that SETs exacerbate (Biddle, 1986; Goode, 1960). Expectancy Theory explained why student motivations distort evaluation results. Students reward instructors who fulfill expectations for leniency, grades, or convenience, and penalize those who do not, meaning that evaluations often reflect satisfaction rather than learning (Uttl & Simbert, 2017; Vroom, 1964). These theories reveal that the flaws in SETs are not incidental but structural, embedded in the very way higher education conceives of accountability, power, and teaching roles.

The ethical implications of relying on flawed SET systems are profound. Faculty are judged by instruments that lack validity, systematically disadvantaging women and minorities and causing significant harm to mental health (Heffernan, 2023b; Lakeman, Coutts, Hutchinson, Massey, Nasrawi, Fielden, et al., 2022). Students, meanwhile, are misled into believing their feedback measures teaching effectiveness, when in fact it is used in ways that reinforce stereotypes and consumerist logics (Marshik et al., 2023). Administrators face ethical challenges of transparency and accountability, as continued reliance on SETs perpetuates injustice under the guise of objectivity. At its core, the reliance on SETs without reform represents an ethical failure to uphold fairness, equity, and integrity in faculty evaluation and student learning.

To address these challenges, this paper proposed several policy reforms. Institutions should adopt a multi-measure framework for teaching evaluation that combines SETs with peer observation, self-reflection, and learning outcome measures (Benton & Cashin, 2012; Berk, 2005). Safeguards must be established to address bias, including disclaimers, filters for abusive comments, and analytic adjustments for systemic disparities (Cunningham et al., 2023). Evaluation frameworks should differentiate by instructional modality, recognizing that online and hybrid teaching require distinct instruments and measures (Zumrawi & Macfadyen, 2023). Institutions must also prioritize faculty well-being and academic integrity by prohibiting the use of abusive comments, supporting faculty development, and reducing incentives for grade inflation(Lakeman et al., 2023). Finally, transparency and accountability in how SET data are used must be enhanced to restore trust and ensure that evaluations align with institutional commitments to equity and rigor (Berk, 2005).

The challenges surrounding SETs are well-documented, yet institutions continue to rely on them because they are convenient, inexpensive, and perceived as legitimate. However, convenience cannot justify practices that perpetuate inequity, distort incentives, and undermine the educational mission. Reform is not optional but essential. Higher education leaders must recognize that SETs, as currently used, measure satisfaction and bias more than they measure teaching quality. Adopting more comprehensive, equitable, and transparent evaluation systems is a matter of justice for faculty, integrity for institutions, and quality for students.

The stakes are clear: if left unreformed, SETs will continue to inflict harm on faculty well-being, perpetuate systemic inequities, and erode the rigor and credibility of higher education. If reformed, however, they can become one component of a more holistic, ethical, and

effective system of evaluating teaching. The choice facing institutions is not whether to change, but how urgently to embrace reform.

References

Benton, S. L., & Cashin, W. E. (2012). Student ratings of teaching: A summary of research and literature. *IDEA Paper, 50*, 1-24.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*(1), 48-62.

Biddle, B. J. (1986). Recent developments in role theory. *Annual Review of Sociology, 12*(1), 67-92. https://doi.org/10.1146/annurev.so.12.080186.000435

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145*, 27-41. https://doi.org/10.1016/j.jpubeco.2016.11.006

Brookfield, S. D. (2005). *The power of critical theory: Liberating adult learning and teaching*. Jossey-Bass.

Buchanan, E. M., Miranda, J. F., & Stephens, C. (2025). The reliability of student evaluations of teaching. *Assessment & Evaluation in Higher Education. Advance online publication*. https://doi.org/10.1080/02602938.2025.2504618

Cunningham, S., Laundon, M., Cathcart, A., Bashar, A., & Nayak, R. (2023). First, do no harm: Automated detection of abusive comments in student evaluation of teaching surveys. *Assessment & Evaluation in Higher Education, 48*(3), 377-389. https://doi.org/10.1080/02602938.2022.2081668

Daskalopoulou, A. (2024). Understanding the impact of biased student evaluations: An intersectional analysis of academics' experiences in the UK higher education context. *Studies in Higher Education, 49*(12), 2411-2422. https://doi.org/10.1080/03075079.2024.2306364

Daumiller, M., Janke, S., Hein, J., Rinas, R., Dickhäuser, O., & Dresel, M. (2023). Teaching quality in higher education: Agreement between teacher self-reports and student evaluations. *European Journal of Psychological Assessment, 39*(3), 176-181. https://doi.org/10.1027/1015-5759/a000700

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review, 48*(2), 147-160. https://doi.org/10.2307/2095101

Fan, Y., Shepherd, L. J., Waters, S. D., Stone, M., Able, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations of teaching: Representation matters. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0209749

Fleming, P. (2019). Reflections on precarity in academia. *Organization Studies, 40*(7), 1305-1317.

Goode, W. (1960). A theory of role strain. *American Sociological Review, 25*(4), 483-496. https://doi.org/10.2307/2092933

Habermas, J. (1984). *The theory of communicative action, Vol. 1: Reason and the rationalization of society*. Beacon Press.

Heffernan, T. (2023a). Abusive comments in student evaluations of courses and teaching: The attacks women and marginalised academics endure. *Higher Education, 85*, 225-239. https://doi.org/10.1007/s10734-022-00831-x

Heffernan, T. (2023b). Abusive student comments in evaluations of teaching: How Australian universities are failing staff. *Higher Education Research & Development, 42*(2), 227-240. https://doi.org/https://doi.org/10.1080/07294360.2021.2002829

Horkheimer, M. (1972). *Critical theory: Selected essays*. Herder and Herder.

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education, 4*(1), 1-8. https://doi.org/10.1080/2331186X.2017.1304016

Johnson, V. E. (2003). *Grade inflation: A crisis in college education*. https://doi.org/10.1007/b97309

Kincheloe, J. L., & McLaren, P. (2002). *Rethinking critical theory and qualitative research. In Y. Zou & E. T. Trueba (Eds.) Ethnography and schools: Qualitative approaches to the study of education*. Rowman & Littlefield.

Lakeman, R., Coutts, R., Hutchinson, M., Massey, D., Nasrawi, D., & Fielden, J. (2022). Appearance, insults, allegations, blame and threats: an analysis of anonymous non-constructive student evaluation of teaching in Australia. *Assessment & Evaluation in Higher Education, 47*(8), 1245-1258. https://doi.org/10.1080/02602938.2021.2012643

Lakeman, R., Coutts, R., Hutchinson, M., Massey, D., Nasrawi, D., Fielden, J., & Lee, M. (2022). Stress, distress, disorder and coping: The impact of anonymous student evaluation of teaching on the health of higher education teachers. *Assessment & Evaluation in Higher Education, 47*(8), 1489-1500. https://doi.org/10.1080/02602938.2022.2060936

Lakeman, R., Coutts, R., Hutchinson, M., Massey, D., Nasrawi, D., Fielden, J., & Lee, M. (2023). Playing the SET game: How teachers view the impact of student evaluation on the experience of teaching and learning. *Assessment & Evaluation in Higher Education, 48*(6), 749-759. https://doi.org/10.1080/02602938.2022.2126430

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*, 291-303. https://doi.org/10.1007/s10755-014-9313-4

Marshik, T., Driscoll, A., A, R., Yu, A., Dutta, N., & Klein, J. (2023). New frontiers in student evaluations of teaching: University efforts to design and test a new instrument for student feedback. *Assessment & Evaluation in Higher Education, 48*(8), 1164-1177. https://doi.org/10.1080/02602938.2023.2190060

Mengel, F., Sauermann, J., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association, 17*(2), 535-566. https://doi.org/10.1093/jeea/jvx057

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology, 83*(2), 340-363. https://doi.org/10.1086/226550

Quansah, F., Adjei, K., & Asante, J. (2024). Validity of student evaluation of teaching in higher education: A systematic review. *Frontiers in Education*, 1-12. https://doi.org/10.3389/feduc.2024.1329734

Sanchez, T., Veas, A., Gilar-Corbi, R., & Castejon, J. (2021). Psychometric perspectives in educational and learning capitals: Development and validation of a scale on student evaluation of teaching in higher education. *Psychological Test and Assessment Modeling, 63*(2), 149-167.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research, 83*(4), 598-642. https://doi.org/10.3102/0034654313496870

Stark, P., & Freishtat, T. (2014). An evaluation of course evaluations. *ScienceOpen Research*, 1-7. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Turner, R. H. (2001). Role Theory. In *Handbook of sociological theory* (pp. 233-254). Springer. https://doi.org/10.1007/0-387-36274-6_11

Uttl, B., & Simbert, D. (2017). Student evaluations of teaching: Teaching quantitative courses can be hazardous to one's career. *PeerJ*. https://doi.org/10.7717/peerj.3299

Van Eerde, W., & Thierry, H. (1996). Vroom's expectancy models and work-related criteria: A meta-analysis. *Journal of Applied Psychology, 81*(5), 575-586. https://doi.org/10.1037/0021-9010.81.5.575

Vroom, V. H. (1964). *Work and motivation*. Wiley.

Wang, H., Ehianeta, R., & Doong, H. (2023). Students' online evaluation of teaching and system continuance usage intention: New directions from a multidisciplinary perspective. *Assessment & Evaluation in Higher Education, 48*(8), 1295-1309. https://doi.org/10.1080/02602938.2023.2199181

Young, M., & McCaslin, S. (2013). Student evaluations of faculty revisited: Online versus traditional methods. *Journal of Academic Administration in Higher Education, 9*(2), 9-17.

Zheng, X., Vastrad, S., He, S., & Ni, C. (2023). Contextualizing gender disparities in online teaching evaluations for professors. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0282704

Zumrawi, A. A., & Macfadyen, L. P. (2023). Proposed metrics for summarizing student evaluation of teaching data from balanced Likert scale surveys. *Cogent Education, 10*(2), 1-13. https://doi.org/10.1080/2331186X.2023.2254665