MARYWOOD UNIVERSITY

STRATEGIC LEADERSHIP & ADMINISTRATIVE STUDIES

Instructional Modality and Student Evaluations of Teaching: A Quantitative Study of Online and Face-to-Face Undergraduate Courses

by

C. Dallas Pierce

A Dissertation in Strategic Leadership & Administrative Studies

Submitted in Partial Fulfillment

of the Requirements for the Degree of

Ph.D. in Strategic Leadership & Administrative Studies

April 2, 2026
Date of Approval

Committee Chair, Dr. Alan Levine

Committee Member, Dr. Jim Mirabella

Patrick Laverty
Patrick Laverty (Apr 3, 2026 14:19:12 MDT)

Committee Member, Dr. Pat Laverty

Carol Kurkowski
Carol Kurkowski (Apr 3, 2026 16:30:15 MDT)

Reader, Dr. Carol Kurkowski

Cristal L. Breitbeil
Cristal L. Breitbeil (Apr 3, 2026 16:03:33 MDT)

Reader, Dr. Cristal Breitbeil

Acknowledgments

I could not have completed this journey without the provision and gentle, loving, and at times loud nudges from my Heavenly Father. All glory belongs to God.

I extend my sincere gratitude to my dissertation chair, Dr. Levine. His support was instrumental in my success. His guidance, encouragement, belief in my lofty goals, and consistent check-ins kept me focused and motivated throughout this process.

I would also like to thank my committee members, Dr. Jim Mirabella and Dr. Pat Laverty. Each of you brought thoughtful insight, valuable feedback, and steady encouragement. Your perspectives strengthened this work in meaningful ways.

My appreciation also goes to my readers, Dr. Carol Kurkowski and Dr. Cristal Breitbeil. Thank you for your time, expertise, and careful review of this dissertation.

I wish my dad could have seen this moment. He left us far too soon, but I know he is in heaven, proudly telling everyone that his daughter is now a doctor. His love and influence continue to guide me every day.

To my family, loved ones, and my closest friends, thank you for your constant support. I see and feel you celebrating alongside me, and that is true love. Beyond prayer, there is nothing more powerful than a group of people who stand beside you through both the challenges and the wins.

To my children, Victoria and Alex, you have always been my motivation to strive for what can feel like unachievable heights. I am deeply grateful for each of you.

With heartfelt gratitude, I thank my entire tribe for the love and support that carried me through this journey.

Table of Contents

Table of Figures

Table of Appendices

Table of Figures

Abstract

Student evaluations of teaching (SET) are widely used in higher education to assess instructional effectiveness and inform personnel decisions. However, questions remain about the extent to which contextual factors, such as instructional modality, influence evaluation outcomes. The purpose of this quantitative study was to examine differences in SET scores between online and face to face undergraduate courses at a community college in Colorado. The dataset included 3,611 course sections evaluated using a standardized 12 item SET instrument over a two year period. Descriptive statistics and correlation analyses were conducted to examine SET patterns, and an independent samples t test was used to compare evaluation scores across instructional modalities. Results indicated a statistically significant difference in SET scores, $t(3609) = -6.88$, $p < .001$, with face to face sections receiving slightly higher ratings than online sections. However, the effect size was small (Cohen's $d = 0.23$), suggesting limited practical differences. These findings contribute to ongoing discussions regarding the interpretation and use of SET in higher education.

Keywords: student evaluations of teaching; instructional modality; online learning; face-to-face instruction; higher education; teaching effectiveness

Chapter 1

The Problem and Its Setting

Introduction

Across higher education institutions globally, student evaluations of teaching (SET) remain one of the most prevalent tools for assessing instructional effectiveness. For decades, universities have relied on these instruments to gather student feedback on various teaching dimensions, with results frequently influencing faculty tenure, promotion, and contract renewals (Spooren et al., 2013; Yamada & Sato, 2021). However, a growing body of research has challenged the validity and reliability of SET, raising concerns about their utility and fairness, particularly in high-stakes evaluative contexts (Hornstein, 2017; Quansah et al., 2024; Spooren et al., 2013; Stroebe, 2020a). Increasingly, studies have highlighted the susceptibility of SET scores to extraneous variables that are unrelated to teaching effectiveness, such as student grades, instructor gender, race, accent, and even physical attractiveness (M. Daskalopoulou, 2024; MacNell et al., 2015; Mengel et al., 2019). Moreover, evidence demonstrates systematic bias against women, faculty of color, and non-native English speakers, raising serious concerns about equity and fairness in their use for high-stakes decisions such as tenure and promotion (M. Daskalopoulou, 2024; T. A. Heffernan, 2022; MacNell et al., 2015; Mengel et al., 2019).

One emerging dimension of this issue relates to instructional modality. Since the shift toward online education, particularly accelerated by the COVID-19 pandemic, universities have expanded their offerings in hybrid and fully online formats. Despite this, institutional reliance on SET results has persisted, with little differentiation between modalities. That is, scores from online and face-to-face (F2F) courses are often treated as equivalent, even though research suggests that student perceptions and rating behaviors may differ across delivery modes (Benton

& Li, 2015a; R. Johnson, 2020). While some studies have explored general trends in SET across modalities, few have controlled for confounding variables by examining SET responses to the same course, across different instructional modalities such as face-to-face and online formats. Addressing this gap is critical to understanding whether instructional modality systematically biases SET results, thereby potentially affecting faculty evaluations in inequitable ways.

SETs are typically administered at the end of academic terms and solicit student perceptions of various aspects of teaching, including clarity, engagement, feedback, and overall effectiveness. Historically, SET instruments have been rooted in psychometric principles, with efforts to validate their structure and interpretation (Marsh, 1984). Nevertheless, critiques have persisted. Scholars have identified numerous validity threats, including halo effects, response biases, and the influence of students' expected grades (Spooren et al., 2013; Uttl et al., 2017). The reliability of SET results has also been questioned, especially when response rates are low or when small sample sizes produce exaggerated statistical variance (Stroebe, 2020a). Further complicating the SET landscape is the impact of instructor demographics. Studies reveal that women and faculty of color receive systematically lower ratings than their white, male counterparts, even when controlling for instructional quality (Bavishi et al., 2010; T. A. Heffernan, 2022; MacNell et al., 2015). These findings raise equity concerns, particularly when SET scores are used for employment-related decisions.

Moreover, the rise of online instruction introduces new complexities. Online students may interpret survey items differently, and the lack of in-person rapport may influence their evaluative tone. Studies suggest that online SETs may yield lower ratings on average or may reflect different dimensions of perceived effectiveness (Benton & Li, 2015a; Chen & Schunn, 2025). Despite this, institutional practices rarely adjust for modality-related effects.

Student evaluations of teaching (SET) are among the most widely used instruments for assessing instructional effectiveness in higher education, yet their validity and reliability remain highly contested (Hornstein, 2017; Quansah et al., 2024; Spooren et al., 2013). Research has consistently shown that SET results are influenced by factors unrelated to teaching quality, including student satisfaction, grading leniency, and contextual variables (Stroebe, 2020a). Moreover, evidence demonstrates systematic bias against women, faculty of color, and non-native English speakers, raising serious concerns about equity and fairness in their use for high-stakes decisions such as tenure and promotion (M. Daskalopoulou, 2024; T. A. Heffernan, 2022; MacNell et al., 2015; Mengel et al., 2019).

A particularly underexamined dimension of this problem is the role of instructional modality. As online education has expanded, especially following the COVID-19 pandemic, institutions have increasingly treated SET results from online and face-to-face courses as equivalent, despite evidence that modality influences student ratings (Benton & Cashin, 2012; Benton & Li, 2015a; R. Johnson, 2020). What remains unclear, however, is whether student responses to SET differ in a course taught in different modalities. This gap is significant, as understanding how modality affects SET scores is essential for ensuring that institutions use evaluations fairly and equitably in faculty assessment.

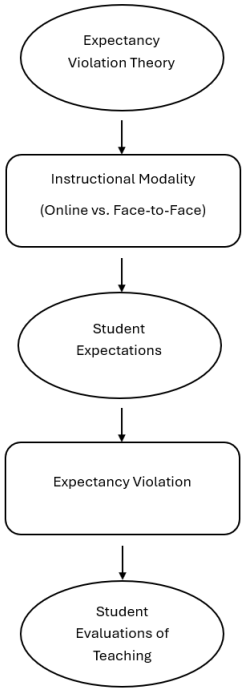
Theoretical Framework

This study was guided by two primary theoretical frameworks: expectancy violation theory (Burgoon, 1993) and social role theory (Eagly, 1987). Expectancy violation theory posits that when student expectations are not met, such as expecting different teaching behaviors in online versus face-to-face settings, they are likely to respond negatively in evaluations. See Figure 1. Expectancy Violation Theory (EVT) proposes that individuals develop expectations for

behavior based on prior experience, context, and social norms (Burgoon, 1993). When reality deviates from these expectations, individuals experience an expectancy violation, which may elicit either positive or negative evaluative responses depending on how the violation is interpreted (Burgoon, 1993). Applied to higher education, EVT suggests that students bring to their courses a set of expectations for instructor immediacy, engagement, and feedback, expectations that are often shaped by prior face-to-face experiences. When instruction occurs in an online environment, reduced visual and nonverbal cues, delayed feedback, or less perceived presence may violate those expectations. Such expectancy violations can lead to lower student evaluations, even when actual instructional quality remains constant.

In the context of SET, this framework helps explain how misalignments between instructional modality and student expectations might influence ratings.

Figure 1 *Expectancy Violation Theory*

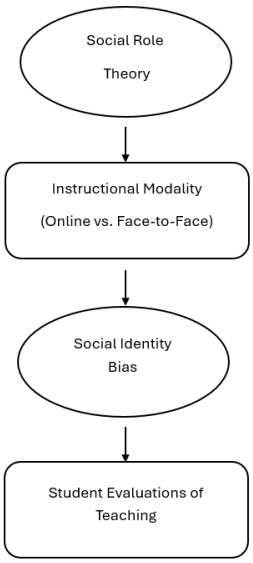


Note: As conceived by researcher

Meanwhile, social role theory highlights how societal stereotypes and expectations related to gender and race can affect perceptions of authority and competence. See Figure 2. Social Role Theory (SRT) complements EVT by explaining how social norms and identity expectations influence evaluative behavior (Eagly, 1987). According to SRT, individuals rely on social scripts and stereotypes to interpret behavior, particularly under conditions of limited information (Eagly, 1987). Within the SET context, this theory helps explain how students' evaluations may reflect not only pedagogical effectiveness but also ingrained expectations about instructor demeanor, communication style, or perceived authority. These role expectations may differ across modalities, for example, online instruction may reduce traditional cues of authority or warmth, prompting students to rely more heavily on assumptions or implicit biases when forming judgments.

This framework provides a lens through which to interpret potential biases in student evaluations, particularly as they intersect with modality.

Figure 2 *Social Role Theory*



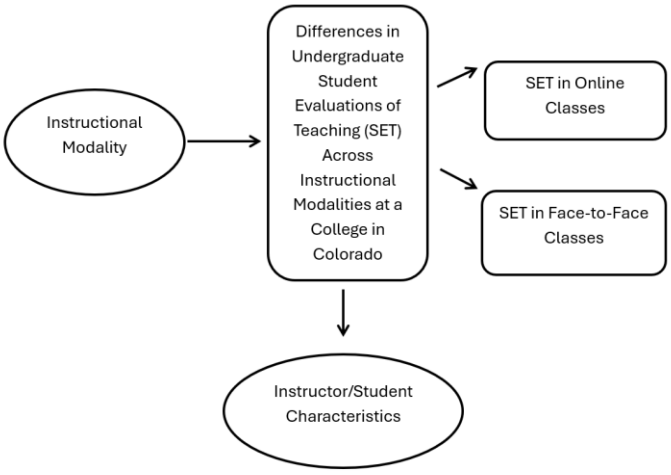
Note: As conceived by researcher

Conceptual Framework

The conceptual framework for this study is grounded in Expectancy Violation Theory (EVT) (Burgoon, 1993) and Social Role Theory (SRT) (Eagly, 1987). These theories collectively explain how students form perceptions of teaching effectiveness and why differences in instructional modality may influence their evaluations.

The conceptual framework (see Figure 1) illustrates how these theoretical lenses interact within the study's variables. Instructional modality (online versus face-to-face) is the independent variable, influencing student evaluation of teaching scores (the dependent variable). EVT explains this relationship through the mechanism of expectation fulfillment or violation, while SRT accounts for potential differences in perception rooted in social and identity-based expectations. Together, the theories provide a lens for interpreting why students may evaluate the same instructor differently across modalities, highlighting how perceptual and social processes shape reported teaching effectiveness.

Figure 3 *Expectancy Violation and Social Role Framework for Student Evaluations of Teaching*



Note: As conceived by researcher

This conceptual framework anchors the study's quantitative, comparative design. It proposes that observed differences in SET scores between online and face-to-face courses stem not only from pedagogical variation but also from how modality affects student expectations and interpretive frameworks. By linking theoretical constructs with empirical measurement, this framework ensures coherence between the research question, *What are the differences in undergraduate student evaluations of teaching across instructional modalities at a college in Colorado?,* and the analytic approach, establishing a foundation for hypothesis testing and interpretation.

Purpose of the Study

The purpose of this non-experimental, comparative survey study was to examine differences in student evaluations of teaching (SET) across two instructional modalities, i.e., online and face-to-face sections, at a college in Colorado. At this stage in the research, student evaluations of teaching (SET), are generally defined as institutional survey instruments designed to capture student perceptions of instructional effectiveness, including engagement, clarity, organization, feedback, and the overall learning environment, administered at the end of each course across modalities.

Research Question

What are the differences in undergraduate student evaluations of teaching (SET) of faculty across instructional modalities, i.e., online and face-to-face sections at a college in Colorado?

Subproblems

Sub-Problem 1, *What are the SET scores in online classes*, was analyzed using a frequency distribution and other descriptive statistics.

Sub-Problem 2, *What are the SET scores in face-to-face classes*, was analyzed using a frequency distribution and other descriptive statistics.

Sub-Problem 3, *What are the differences in SET scores across modalities*, was analyzed using an independent samples t-test.

Hypotheses

H_0: There are no significant differences in SET item scores of faculty across instructional modalities, i.e. online and face-to-face classes at a college in Colorado

H_a: There are significant differences in SET item scores of faculty across instructional modalities, i.e. online and face-to-face classes at a college in Colorado

Definition of Terms

Undergraduate: Undergraduate education refers to the foundational stage of postsecondary study undertaken by students pursuing a bachelor's degree or equivalent qualification, typically representing the first level of tertiary education following secondary school (Papadogiannis et al., 2024; UNESCO, 2023). For this study, undergraduate designates courses classified by the institution's registrar as part of an associates or bachelor's-level program.

Student Evaluations of Teaching (SET): Institutional survey instruments used to assess students' perceptions of teaching effectiveness, typically administered at the end of a course (Hornstein, 2017; Spooren et al., 2013). In this study, SET refers to the college's end-of-term evaluation instrument, which includes Likert-scale questions and open-ended comments. Data was be used at the course-section level for comparative analysis.

Faculty: Individuals whose official appointment is primarily to teach (or both teach and research) at the professional level, regardless of their full-time, part-time, adjunct, or tenure-track status (Monnier, May-June 2017). In this study, "faculty" refers to all full-time and part-time faculty (tenure-track, non-tenure-track, adjunct) who taught the course sections in the dataset.

Instructional Modality: The format through which instruction is delivered, specifically face-to-face (F2F) or online (asynchronous or synchronous) (Benton & Li, 2015a). For this study, instructional modality is coded by registrar designation as face-to-face, online-asynchronous, and serves as a categorical variable in the analysis.

Online Classes: Online classes are courses in which instruction and interaction occur primarily through digital platforms, either synchronously in real time or asynchronously through recorded materials and discussion forums (Arias et al., 2018; Tu & Adkins, 2020). For this study, online classes are coded by registrar designation as online-asynchronous, where students complete coursework independently through the learning management system.

Face-to-Face Classes: Face-to-face classes involve direct, in-person interaction between instructors and students in a shared physical classroom at scheduled times (Arias et al., 2018; Tu & Adkins, 2020). For this study, face-to-face classes are coded by registrar designation as in-person courses where instructional contact hours occur on campus in real time.

College: A college is an accredited institution of higher education that offers undergraduate programs leading to associate or bachelor's degrees and, in some cases, limited graduate study. Colleges emphasize teaching and learning, focusing on academic instruction and student development rather than extensive research or professional specialization (Carnegie Classification of Institutions of Higher Education, 2022; U.S. Department of Education, 2023). For this study, college refers to the state institution where the data are collected, an accredited public higher education institution offering associate and bachelor's degree programs in Colorado.

Delimitations

This study was delimited to undergraduate courses taught by instructors and faculty at a midsized college in Colorado. Only courses taught in both face-to-face and online modalities within the same academic year were included. The study focused exclusively on quantitative SET data collected through institutional end-of-course evaluations.

Assumptions

It is assumed that students respond to SET instruments honestly and to the best of their ability, in an unbiased manner. It is also assumed that the faculty deliver equivalent content and use comparable pedagogical strategies across modalities. Lastly, it is assumed that institutional SET instruments are valid measures of instructional quality, at least in terms of capturing student perceptions.

Significance of the Study

This study addresses several gaps in the literature on SET, most notably the comparative evaluation of teaching effectiveness across modalities. Its findings may help inform institutional policies on how to interpret and use SET results fairly, especially in high-stakes decisions. It may also guide the development of more nuanced SET instruments that account for modality effects and minimize potential biases. Furthermore, by integrating theoretical perspectives on expectancy and social role, the study advances the conceptual understanding of how students form evaluative judgments in diverse instructional settings. This research holds practical significance for academic leaders, policymakers, faculty development specialists, and faculty themselves.

This study contributes to the ongoing discourse on the validity and fairness of student evaluations of teaching (SET) by examining whether instructional modality introduces

systematic differences in ratings. Despite substantial growth in online learning, few empirical studies have directly compared SET results across modalities (Benton & Li, 2015a; Rivera & Beranek, 2023). By focusing on modality comparisons, this research isolates modality effects from instructor characteristics and course content, providing clearer evidence of whether differences in SET scores represent true variations in instructional quality or contextual bias.

The findings have direct implications for institutional policy and faculty evaluation practices. If SET results are shown to vary significantly by modality, colleges must reconsider the equity of using unadjusted SET averages in high-stakes decisions such as tenure, promotion, and merit pay (Hornstein, 2017; Linse, 2017). Such evidence can inform the development of modality-adjusted benchmarks or contextual weighting systems that ensure evaluations are interpreted fairly. Moreover, this study supports the integration of multi-source assessment frameworks, including peer observation and teaching portfolios, thereby aligning institutional evaluation processes with principles of validity, reliability, and nonmaleficence (Crimmins et al., 2024).

From a theoretical perspective, the study expands the application of Expectancy Violation Theory (Burgoon, 1993) and Social Role Theory (Eagly, 1987) within the domain of teaching evaluation. Expectancy Violation Theory helps explain how students' preconceived expectations of engagement and immediacy differ between online and face-to-face modalities while Social Role Theory provides a lens for understanding how gendered and racialized expectations may amplify evaluation bias in technology-mediated environments. Together, these frameworks deepen understanding of how perceptual and social dynamics shape evaluative behavior.

Finally, this study has practical significance for faculty development and instructional design. Insights from this research may guide faculty in adapting communication, feedback, and

engagement strategies to meet students' modality-specific expectations, potentially improving both student satisfaction and learning outcomes. By clarifying how modality interacts with student perception, this study offers actionable evidence for creating more equitable, evidence-based, and contextually sensitive evaluation systems in higher education.

Chapter 2

Literature Review

Introduction and Historical Context

Student evaluations of teaching (SET) have long served as a standard mechanism for capturing student perceptions of instructional quality in higher education. These evaluations are typically administered at the end of an academic term and are used to inform decisions related to faculty performance, promotion, and curriculum development (Spooren et al., 2013). Their popularity stems from their cost-effectiveness, ease of administration, and ability to generate large volumes of data. However, despite their widespread use, SETs have been the subject of considerable debate for decades regarding their validity, reliability, and potential for bias.

The origins of SETs in higher education can be traced to the mid-20th century, when institutions began seeking more structured feedback mechanisms to assess teaching effectiveness in response to growing demands for accountability and quality assurance (Berk, 2005). Early iterations of SET instruments focused on general impressions of teaching, but over time, they evolved to include more specific indicators such as clarity, engagement, organization, and perceived learning outcomes (Marsh, 1980). By the 1990s and early 2000s, SETs had become deeply embedded in institutional evaluation processes across North America, Europe, and Asia (Beran & Rokosh, 2009).

Despite their institutional entrenchment, critiques of SETs emerged early and have intensified over the past two decades. Scholars have repeatedly questioned whether student perceptions accurately reflect instructional quality. Research by Marsh and Roche (1997) supported the multidimensionality of teaching effectiveness and argued that students are capable of evaluating distinct aspects of instruction. However, subsequent studies began to challenge this

position. For instance, Uttl et al. (2017) found no meaningful correlation between SET scores and student learning outcomes, casting doubt on the assumption that higher ratings necessarily indicate better teaching.

As SETs became integral to faculty evaluation and institutional rankings, concerns about their fairness and equity grew. The potential for SETs to reinforce biases, particularly against women, faculty of color, and non-native English speakers, has been documented in numerous studies (Basow & Martin, 2012; Kreitzer & Sweet-Cushman, 2021). In addition, non-pedagogical factors such as course difficulty, grading leniency, and even instructor attractiveness have been shown to influence ratings (Boring et al., 2016).

The COVID-19 pandemic and the subsequent rise of remote learning reignited debates about the appropriateness of SETs in evaluating instructional effectiveness. With courses rapidly shifting to online formats, institutions continued to rely on SETs despite limited evidence about their comparability across modalities. This situation prompted renewed scrutiny of how different delivery formats might impact student perceptions and evaluation patterns (M. Johnson, 2020).

Recent literature has shifted from asking whether SETs should be used to how they can be used responsibly. Researchers and institutional leaders alike have emphasized the importance of contextualizing SET data within broader evaluation frameworks, including peer reviews, self-assessments, and learning analytics (Linse, 2017). Emerging studies call for a reexamination of the psychometric properties of SET instruments and advocate for instruments better aligned with actual teaching behaviors and student learning (Chen & Schunn, 2025; Quansah et al., 2024).

The Evolution and Purpose of Student Evaluations of Teaching (SET)

Student evaluations of teaching (SET) emerged in the mid-twentieth century as part of a larger movement toward evidence-based assessment of instruction in higher education. Early

iterations of student feedback instruments were designed primarily as formative tools, mechanisms through which instructors could obtain structured impressions of their teaching methods and make improvements (Marsh, 1980). The original intent aligned closely with the principles of continuous quality improvement: students, as direct participants in the learning process, were viewed as uniquely positioned to comment on course clarity, organization, and engagement (Benton & Cashin, 2012). Institutions in the United States adopted paper-based surveys that focused on observable behaviors such as clarity of presentation or preparedness, and the aggregated results were shared privately with instructors for self-reflection.

By the late 1960s and 1970s, the spread of standardized questionnaires across universities reflected the growing professionalization of teaching. Researchers sought to establish whether student ratings were valid indicators of teaching quality and to what extent they correlated with student learning outcomes (Cohen, 1981; Marsh, 1980). Meta-analyses during this period reported moderate positive relationships between ratings and achievement, reinforcing confidence in the tool's utility. Yet even at that stage, scholars cautioned against over-interpreting SET data, noting that ratings could be influenced by variables such as expected grades, course difficulty, or student motivation (Costin et al., 1971). Consequently, SET were recommended as one source of feedback among several, to be triangulated with peer observation and teaching portfolios.

The purpose of SET began to shift significantly in the late twentieth century as universities confronted external pressures for accountability and transparency. Administrative interest in quantifiable metrics grew, and SET results were gradually incorporated into personnel decisions involving tenure, promotion, and merit pay (Spooren et al., 2013). This transformation changed both the stakes and the meaning of student feedback. What had once been an internal

developmental exercise became a public performance measure. Stark and Freishtat (2014) argue that this managerial use of SET distorted its formative intent by equating popularity with pedagogical effectiveness. The shift also created incentives for faculty to pursue higher ratings through lenient grading or reduced rigor, behaviors that potentially undermine learning quality (Stroebe, 2020a).

As digital technologies advanced, institutions replaced paper forms with online survey platforms, citing efficiency and sustainability benefits. This modernization expanded reach but introduced new methodological complications. Response rates declined sharply when surveys moved online, often dropping 20–30 percentage points compared to in-class completion (Benton & Li, 2015a). The resulting non-response bias weakened representativeness and raised doubts about whether numerical averages accurately reflected collective student opinion. Moreover, anonymity in online systems, initially intended to encourage honesty, sometimes facilitated unconstructive or even abusive comments, creating emotional strain for instructors (Heffernan, 2023; Lakeman, Coutts, Hutchinson, Massey, Nasrawi, & Fielden, 2022).

The COVID-19 pandemic further accelerated the evolution of SET, forcing institutions to evaluate teaching under unprecedented conditions of disruption. As instruction moved almost entirely online, traditional assumptions about classroom engagement and instructor presence no longer applied. Crimmins et al. (2024) observed that universities' continued reliance on SET during the pandemic amplified ethical tensions, as instruments intended for formative feedback were deployed summatively in a period of widespread pedagogical uncertainty. Similarly, Sidwell et al. (2025) found that faculty reported heightened stress and diminished trust in evaluation systems during this time, highlighting the emotional toll of student feedback collected under crisis conditions. These findings suggest that the post-pandemic period marked a turning

point toward reexamining SET practices, with renewed calls for ethical, contextualized, and well-being–oriented interpretation (Macfayden et al., 2023).

Over the past two decades, theoretical debates surrounding SET have expanded beyond questions of psychometric validity to encompass cultural and ethical dimensions. Hornstein (2017) contends that universities' continued reliance on SET demonstrates an institutional contradiction: while academic rigor is emphasized in scholarship, decisions about teaching quality often rest on instruments lacking comparable scientific rigor. Similarly, Spooren et al. (2013) identify a "misalignment of purpose," wherein surveys designed for formative reflection are deployed summatively for evaluation and ranking. This dual use not only conflates distinct purposes but also alters student behavior, as learners aware of the evaluative consequences may view the survey as an opportunity to reward or penalize instructors rather than to provide constructive feedback.

Another strand of scholarship links the evolution of SET to broader societal trends in consumerism within higher education. As tuition costs increased and institutional marketing adopted customer-service language, students began to perceive themselves as consumers whose satisfaction deserved formal measurement (Boring, 2017; Stark & Freishtat, 2014). The design of many SET instruments reflects this orientation, emphasizing enjoyment, clarity, and perceived fairness over cognitive gains or long-term learning (Boring, 2017; Stark & Freishtat, 2014). Critics argue that this "customer satisfaction" model erodes the academic ethos of shared responsibility for learning and frames education as a transactional service. From this perspective, SET serve not only as a feedback tool but also as a symbolic reassurance to stakeholders that institutions listen to student voices.

The institutionalization of SET is not confined to the United States. As Pineda and Steinhardt (2023) note, the global spread of standardized teaching evaluations reflects a process of institutional convergence across higher education systems. Through mechanisms of normative, coercive, and mimetic isomorphism (DiMaggio & Powell, 1983), universities have adopted similar instruments to signal accountability and quality assurance to external stakeholders. This international diffusion underscores the symbolic rather than empirical role of SET in many contexts, functioning as an emblem of modern governance and institutional legitimacy even where evidence of validity remains contested.

Despite persistent critiques, SET remain ubiquitous worldwide. Their endurance is explained partly by administrative convenience, data are inexpensive to collect and easy to quantify, and partly by the absence of equally scalable alternatives (Spooren et al., 2013). Advocates continue to claim that, when carefully interpreted, SET provide valuable insights into student perceptions that can complement peer review and learning assessments (Benton & Li, 2015b). Yet even proponents concede that interpretation must occur within contextual boundaries, accounting for discipline, course level, and instructor experience.

Recent developments in educational technology point toward yet another phase in the evolution of SET. Emerging tools such as automated sentiment analysis of open-ended comments (Cunningham et al., 2023) and AI-driven evaluation dashboards (Wang & Zhao, 2025) promise greater efficiency and scalability. However, scholars caution that these data-intensive systems may inadvertently replicate existing biases at a larger scale, reinforcing earlier critiques of SET as a superficial or inequitable proxy for teaching quality. This technological turn illustrates both the adaptability and the persistent challenges of SET as institutions continue to modernize their evaluation practices.

In sum, the historical trajectory of SET reflects a steady drift from formative feedback toward summative judgment. What began as an instrument to improve teaching practice has evolved into an institutional mechanism for accountability and comparison. This evolution has generated enduring tensions: between development and control, between learning outcomes and satisfaction, and between scholarly ideals of evidence and administrative demands for simplicity. Understanding this trajectory provides a necessary foundation for examining contemporary concerns about validity, bias, and modality differences explored in subsequent sections of this literature review.

Validity, Reliability, and Psychometric Issues

The credibility of student evaluations of teaching (SET) rests on two interrelated assumptions: that students are capable of reliably assessing teaching effectiveness, and that the instruments used are valid measures of instructional quality. These assumptions have been the subject of sustained empirical scrutiny. While SETs have demonstrated moderate reliability in capturing students' perceptions of the classroom experience, their validity as measures of teaching quality remains contested (Spooren et al., 2013; Uttl et al., 2017).

Recent systematic reviews confirm that the validity of SET remains inconsistent and context-dependent. Quansah et al. (2024) analyzed 98 empirical studies across 22 countries and concluded that SET validity depends heavily on contextual variables such as grading leniency, course difficulty, and class size. Similarly, Chen and Schunn (2025) demonstrated that even when SET items display high internal reliability, they often fail to predict measurable learning outcomes or motivation. These findings echo Uttl and Simbert (2017), who found little correlation between SET scores and student achievement, reinforcing the view that evaluations measure perceived satisfaction rather than instructional quality.

Validity refers to the degree to which an instrument measures what it purports to measure. Several recent reviews have questioned the construct validity of SET instruments. A systematic review by Quansah et al. (2024) identified multiple sources of measurement error in SET administration, including ambiguous item wording, inconsistent scales, and lack of theoretical grounding in instrument design. The authors concluded that while SETs are often reliable in terms of internal consistency, they fall short in establishing content and criterion-related validity, particularly when used to infer teaching effectiveness.

Using confirmatory factor analysis, Senden et al. (2023) validated a three-factor model of teaching quality, classroom management, cognitive activation, and supportive climate, using international teaching data. Their results suggest that while SET instruments capture structural aspects such as organization and clarity, they are weaker indicators of pedagogical depth or learning impact. This aligns with the multidimensional frameworks proposed by Marsh (1984, 1987), but current evidence indicates overlapping constructs and limited discriminant validity across items, raising concerns about how well SET scores isolate true teaching effectiveness.

The reliability of SETs has also been a focus of recent empirical work. For example, Beránek and Zvára (2022) conducted a longitudinal psychometric analysis across three universities and found that SET scores remained stable across multiple administrations, suggesting acceptable temporal reliability. However, they also observed substantial within-course variability that could not be accounted for by changes in teaching practices, raising concerns about potential halo effects and context-specific biases.

Psychometric analyses have increasingly adopted factor analytic techniques to evaluate the internal structure of SET instruments. A study by Chen and Schunn (2025) found that commonly used SET scales often include latent constructs that do not align with theoretical

models of instructional effectiveness. Their findings emphasized the need for more rigorous construct validation procedures, particularly for instruments used in high-stakes evaluation settings.

Contemporary researchers have adopted advanced psychometric models to evaluate the stability of SET scores. Kember et al. (2023) used Generalizability Theory (GT) to partition variance attributable to students, instructors, and courses, revealing that non-instructional factors explain a large share of score variability. Similarly, Fynn (2024) applied Item Response Theory (IRT) and Many-Facet Rasch Modeling (MFRM) to identify rater severity and modality-related bias. These approaches offer a more comprehensive understanding of reliability than single-coefficient measures such as Cronbach's alpha, helping to clarify whether differences in SET truly reflect instructional quality or contextual noise.

Moreover, the dimensionality of teaching quality remains a complex and unresolved issue. Traditional SET instruments tend to aggregate diverse teaching behaviors, such as clarity, engagement, and feedback, into a single global score. This practice has been criticized for masking the multidimensional nature of teaching and for failing to differentiate between pedagogical domains that may vary in importance depending on the course context (Linse, 2017; Marsh, 1987).

Alternative approaches to SET design are beginning to emerge. For instance, Fynn (2024) recommends integrating indicators from learning sciences and instructional design theory to ensure that instruments capture pedagogical strategies that are empirically linked to student achievement. Similarly, Quansah et al. (2024) call for participatory instrument development, wherein faculty and students collaboratively identify relevant evaluation domains, enhancing both content validity and user acceptance.

Emerging technologies are being used to analyze SET data in new ways. For example, Cunningham et al. (2023) employed AI-based sentiment analysis to evaluate tone and emotional valence in open-ended comments, while Wang and Zhao (2025) introduced a fuzzy-logic evaluation framework that applies weighting algorithms to linguistic data. Although these models aim to improve objectivity and consistency, researchers warn that algorithmic decision systems may reproduce existing social biases and magnify inequities if not carefully monitored (T. A. Heffernan, 2022). These innovations illustrate both the promise and pitfalls of data-driven evaluation methods in higher education.

Finally, several scholars have noted the growing disconnect between what SETs measure and what institutions aim to improve. Linse (2017) and Uttl et al. (2017) argue that without psychometric rigor, SETs risk perpetuating superficial metrics of popularity rather than promoting genuine instructional improvement. This misalignment can lead to unintended consequences, such as grade inflation, strategic self-presentation by instructors, and a reluctance to adopt innovative or challenging pedagogical approaches.

In summary, the psychometric evaluation of SET remains a contested field. While classical test theory provided foundational support for reliability, modern frameworks such as Generalizability Theory and Rasch modeling reveal that contextual factors significantly distort score interpretation. Furthermore, limited evidence links SET scores to genuine learning outcomes, leaving questions about construct validity unresolved. These gaps justify ongoing comparative studies that examine how SET operates across instructional modalities, particularly whether online and face-to-face environments produce systematically different patterns of evaluation.

Instructional Modality and SET Comparisons

One of the most significant yet underexplored variables in the interpretation of student evaluations of teaching (SET) is instructional modality. The transition to online education, accelerated by the COVID-19 pandemic, has prompted scholars to question whether SET scores are directly comparable between online and face-to-face courses. Existing research suggests that modality may influence SET scores in both subtle and substantial ways, raising concerns about fairness and interpretability (M. Johnson, 2020).

Recent comparative studies have examined whether instructional modality influences SET outcomes under controlled conditions. Aldeman et al. (2025) compared identical courses taught online and face-to-face by the same instructors, finding significantly lower student ratings in online-synchronous and asynchronous sections. Qin (2025) reported similar results over a three-year longitudinal design, concluding that while student learning outcomes were equivalent across modalities, perceived instructional effectiveness was consistently higher in face-to-face classes. Tu and Adkins (2020) also found that students rated face-to-face instruction most favorably, particularly for communication, engagement, and clarity, whereas asynchronous online courses received the lowest scores. These findings suggest that modality continues to exert a measurable influence on SET outcomes even when course content and instructors remain constant.

A recent meta-analysis by Rivera and Beranek (2023) reviewed over 60 studies comparing SET scores across instructional modalities and found that approximately half reported significant differences, favoring face-to-face delivery. The authors note that these differences often diminish when factors such as class size, instructor experience, and grading leniency are statistically controlled. Similarly, longitudinal research (Macfayden et al., 2023) indicates that as

online instruction has become more normalized since the pandemic, evaluation gaps between modalities have narrowed, suggesting that adaptation and digital literacy may mediate student perceptions of teaching effectiveness.

Several comparative studies have examined whether SET ratings differ systematically across modalities when controlling for instructor and course. For instance, Benton and Li (2015b) found that online courses often receive lower scores on dimensions such as instructor availability and engagement, while face-to-face courses tend to be rated more favorably for classroom interaction and responsiveness. However, these differences are not always consistent. Macfayden et al. (2023) noted that online courses may outperform face-to-face courses on criteria such as organization and clarity, especially when instructors have adapted course materials effectively for digital delivery.

More recent studies have used matched-course designs to minimize confounding variables. Bunn et al. (2021) conducted a large-scale study comparing SET scores from over 1,200 paired course sections taught by the same instructors across online and in-person modalities. Their findings showed statistically significant differences in student ratings, particularly in items related to immediacy, feedback, and overall course satisfaction. However, these differences were also mediated by students' prior experiences with online learning and their self-reported digital literacy levels.

Expectancy violation theory (Burgoon, 1993) provides one framework for understanding why online courses often receive lower evaluations. EVT posits that individuals form expectations about interpersonal communication norms, and when those expectations are violated, such as through reduced immediacy, nonverbal cues, or feedback lag, perceived effectiveness declines. In educational settings, students may subconsciously equate physical

presence with engagement, leading to negative expectancy violations in online environments even when instructional quality is comparable.

Social role theory (Eagly, 1987; Goode, 1960) also helps explain differences in evaluation patterns. This theory suggests that gendered and social expectations shape perceptions of competence and authority. When instruction occurs in less traditional, technology-mediated environments, these role expectations can intensify, often to the disadvantage of women and marginalized instructors. Studies such as T. A. Heffernan (2022) and A. Daskalopoulou (2024) confirm that online platforms may amplify bias through disinhibition effects and anonymized feedback mechanisms.

Researchers emphasize that differences in SET across modalities are not purely technological but often mediated by contextual factors. Course level, disciplinary expectations, and class size significantly shape ratings (Benton & Li, 2015b; Rivera & Beranek, 2023). Grading leniency and perceived workload have also been shown to moderate evaluation outcomes, with more rigorous or quantitative courses typically receiving lower scores (Uttl & Simbert, 2017). Additionally, online environments rely heavily on written communication, which may disadvantage instructors who are non-native speakers or who teach abstract, process-oriented subjects (Tu & Adkins, 2020). These findings suggest that modality interacts with social and pedagogical variables, complicating simplistic comparisons of SET means.

The instructional modality also appears to influence students' interpretation of SET items. In a qualitative study, Martin and Borup (2022) found that online learners often interpret engagement and feedback differently than face-to-face students. For example, in an online context, "engagement" may be interpreted as prompt email responses or regular announcements, whereas in-person students may associate it with physical presence and enthusiasm during

lectures. These divergent interpretations complicate the use of uniform SET instruments across modalities.

Another important consideration is the role of instructional design. Online courses that incorporate active learning strategies, structured feedback, and interactive technologies tend to receive more favorable evaluations than those that rely heavily on passive content delivery. A study by Gupta and Henricksen (2024) confirmed that course design, rather than modality alone, is a significant predictor of SET outcomes. Their findings suggest that the pedagogical quality embedded in each modality should be accounted for when comparing SET scores.

Modality differences in SET scores may also reflect deeper structural inequalities. For example, Aragon and Davis (2022) found that online instructors, particularly adjunct faculty, are often evaluated more harshly due to perceptions of reduced legitimacy and engagement. This bias is exacerbated in institutions that do not provide adequate training or support for online instruction.

While many studies have demonstrated significant modality-based differences in SET outcomes, others have reported null findings. According to a meta-analysis by Rivera and Beranek (2023), about 40 percent of comparative studies report no significant differences between online and face-to-face evaluations. This inconsistency may be due to variability in course types, student demographics, and institutional support systems.

Thematic Dimensions of SET

Student evaluations of teaching (SET) instruments often include multiple dimensions designed to capture the multifaceted nature of instructional effectiveness. Commonly assessed dimensions include student engagement, instructor clarity, course organization, timeliness and quality of feedback, and the overall learning environment. While these dimensions are intended

to offer a granular view of teaching quality, their interpretation and reliability can vary significantly across students and instructional contexts.

Contemporary psychometric research has refined the dimensional structure of SET instruments. Senden et al. (2023) identified three consistent latent constructs, classroom management, cognitive activation, and supportive climate, across international samples, mirroring and updating Marsh's earlier model. Similarly, Sanchez et al. (2021) applied item response theory to develop a 32-item SET scale that confirmed distinct but correlated factors for organization, clarity, and feedback timeliness. These studies affirm that SET captures recognizable aspects of instructional practice, yet they also reveal significant overlap between affective (i.e., enthusiasm) and cognitive (i.e., clarity) dimensions, suggesting that students may conflate teaching behaviors with personal attributes.

Engagement is frequently included as a central component of SET instruments, yet it remains one of the most variably interpreted constructs. In face-to-face settings, engagement is often linked to the instructor's expressiveness, ability to stimulate discussion, and presence in the classroom (Clayson, 2021). In contrast, in online settings, students often equate engagement with prompt email responses, regular announcements, and the availability of digital resources (Martin & Borup, 2022). These findings suggest that engagement may not be a unidimensional construct, and its operationalization should be contextually defined by modality.

Clarity and organization are generally considered robust predictors of positive SET scores across modalities. According to Marsh and Roche (1997), clarity of instruction is one of the strongest correlates of students' overall satisfaction. Recent studies have confirmed this trend. In a study by Lewis and Freeman (2023), both online and in-person students rated clarity as the most important dimension influencing their overall evaluation. However, the authors noted

that online students tend to value visual organization of course content (i.e., LMS structure) more than oral clarity, which is typically emphasized in traditional classrooms.

Research indicates that students in different cultural and disciplinary settings emphasize distinct facets of teaching quality. Tomas et al. (2022) found that nursing students in Namibia prioritized instructor approachability and clarity more than organization or feedback speed, while Jabeen et al. (2024) observed that Pakistani higher-education students weighted fairness and enthusiasm highest. In contrast, U.S. studies often highlight clarity, engagement, and perceived care as dominant predictors of high ratings (Benton & Li, 2015b). These cross-context findings underscore that SET dimensions are not universal constructs but socially and institutionally mediated perceptions of effective teaching.

Meta-analytic evidence suggests that clarity and organization remain the most reliable and valid predictors of overall teaching effectiveness scores (Benton & Cashin, 2012; Rivera & Beranek, 2023). In contrast, affective traits such as enthusiasm and friendliness exhibit weaker predictive validity once cognitive factors are controlled (Chen & Schunn, 2025). However, the persistent correlation between affective and cognitive dimensions complicates construct interpretation. As Spooren et al. (2013) note, SET items often reflect students' emotional reactions rather than discrete instructional practices, challenging efforts to use dimensional subscales as objective indicators of performance.

Comparative research reveals that instructional modality can shift the relative weight of SET dimensions. Tu and Adkins (2020) reported that in online courses, students valued prompt feedback and communication clarity more highly than enthusiasm or classroom engagement. Similarly, Aldeman et al. (2025) found that "responsiveness" replaced "approachability" as the most salient factor in online-synchronous settings. Qin (2025) further demonstrated that while

cognitive dimensions such as clarity and structure remained stable across modalities, affective dimensions, such as immediacy and instructor warmth, declined in online environments. These patterns suggest that the meaning of effective teaching may be modality-contingent, reflecting differing communication norms and expectations.

Feedback is another crucial dimension, particularly in formative evaluation processes. Timely and constructive feedback has consistently been associated with higher SET ratings (Gupta & Henricksen, 2024). In online courses, the asynchronous nature of interaction places additional weight on written feedback, which students often interpret as a proxy for instructor involvement and care (Borup et al., 2021). Face-to-face students, on the other hand, may receive informal feedback during class discussions or office hours, which may not always be captured by SET instruments.

The learning environment, encompassing both physical and psychological aspects, also contributes significantly to SET outcomes. A supportive and inclusive environment has been shown to positively impact student perceptions, especially for marginalized groups (Aragon & Davis, 2022). In online courses, learning environment is shaped by factors such as ease of navigation, access to resources, and the tone set by the instructor in communications and discussion boards.

These thematic dimensions are not only central to understanding SET results but also to improving teaching practice. However, their measurement often lacks consistency across instruments and institutions. A review by Quansah et al. (2024) emphasized the need for dimension-specific validation, noting that items intended to measure the same construct often load onto different factors depending on modality and course type.

Furthermore, intercorrelations among SET dimensions can obscure their diagnostic utility. For instance, high scores in clarity often correlate with high scores in organization and engagement, making it difficult to determine which specific behaviors are driving student satisfaction. Researchers have called for the development of more discriminant metrics that can isolate instructional behaviors relevant to each dimension (Chen & Schunn, 2025).

Collectively, these findings illustrate that SET is a multidimensional construct influenced by cultural, disciplinary, and technological contexts. While dimensions such as clarity and organization are consistently valued, others, like engagement, enthusiasm, or fairness, fluctuate across modalities and student populations. This multidimensionality complicates the interpretation of SET scores as singular measures of teaching quality. For the current study, recognizing these dimensions is essential for interpreting whether modality-based differences reflect variations in perceived instructional behaviors or shifts in the weighting of underlying evaluation factors.

Bias and Equity in SET

Bias in student evaluations of teaching (SET) has emerged as one of the most persistent and well-documented concerns in higher education. Numerous studies have revealed that SET instruments are susceptible to implicit and explicit biases related to gender, race, accent, perceived age, and even physical appearance (Basow & Martin, 2012; Fan et al., 2019). These biases not only compromise the validity of SET results but also perpetuate structural inequalities in faculty evaluations, promotion, and retention decisions.

Gender bias is among the most extensively studied forms of bias in SET. Research shows that female instructors are often rated lower than their male counterparts, even when teaching the same content in similar formats. MacNell et al. (2015) demonstrated that students rated

instructors they believed to be male more favorably than those they believed to be female, despite the fact that the same individual taught both courses under pseudonyms. More recent evidence from Boring et al. (2016) reinforces these findings, indicating that gender bias can significantly distort SET outcomes, especially in male-dominated disciplines.

Scholars increasingly emphasize that bias in SET cannot be reduced to gender alone but must be understood through an intersectional lens. A. Daskalopoulou (2024) found that women and faculty of color reported disproportionately higher exposure to disrespectful or abusive comments, especially in anonymous online systems. M. Daskalopoulou (2024) similarly observed that women of color experienced compounded bias in open-text comments that referenced both appearance and perceived authority. Mengel et al. (2019) demonstrated that women received significantly lower ratings in large courses, suggesting that intersectional bias interacts with structural factors like class size and teaching format. Together, these studies reveal that SET bias is cumulative and contextually embedded, rather than uniform across faculty groups. Racial and ethnic biases also distort SET results. Studies by Chavez and Mitchell (2021) and Mengel et al. (2019) found that faculty of color, particularly Black and Latinx instructors, receive systematically lower SET scores than white colleagues. These disparities persist even after controlling for course content, modality, and instructor qualifications. Moreover, such biases are exacerbated when students evaluate instructors from historically marginalized groups in courses dealing with race, gender, or social justice (A. Heffernan, 2022).

Research indicates that online and anonymous evaluation systems amplify bias through disinhibition effects. T. Heffernan (2022) analyzed thousands of open-ended SET comments and found a higher prevalence of personal insults, gendered language, and tone policing in online submissions. Lakeman, Coutts, Hutchinson, Massey, Nasrawi, Fielden, et al. (2022) and

Cunningham et al. (2023) corroborated these findings through sentiment analysis, identifying greater emotional negativity and hostility in online feedback compared to in-person evaluations. A. Daskalopoulou (2024) note that the perceived anonymity of online platforms reduces accountability, allowing stereotypes about gender, ethnicity, and age to manifest more overtly. This evidence suggests that modality not only affects rating patterns but also shapes the tone and content of qualitative feedback.

Linguistic bias is another important dimension, particularly for non-native English speakers. Subtly negative perceptions about accents or grammatical errors can lead students to underrate instructors' competence and communication skills (Zhou & Sloan, 2022). Studies show that instructors who speak with non-native or regional accents are often rated lower on communication and clarity items, even when teaching effectiveness is equivalent (Boring et al., 2016; Uttl & Simbert, 2017). Rivera and Beranek (2023) found that accent-related perceptions were particularly salient in online environments, where vocal inflection and pronunciation substitute for in-person cues such as gesture and facial expression. These biases are especially salient in online modalities, where audio quality and asynchronous communications may amplify misperceptions.

In addition to demographic biases, affective and contextual factors also influence SET results. Stroebe (2020b) argues that students' ratings are frequently shaped by their expectations, grades, and emotional responses to the course, rather than by objective indicators of instructional quality. This "consumer mindset" often penalizes instructors who maintain rigorous academic standards, especially in quantitative or high-demand courses.

Anonymous student feedback, while designed to encourage honesty, can also open the door to hostile or abusive comments. Women and marginalized instructors are disproportionately

targeted by derogatory and non-constructive feedback (Flaherty, 2023; Mitchell & Martin, 2023). These comments not only impact SET scores but also contribute to faculty stress and burnout, undermining the goal of constructive evaluation.

Efforts to mitigate bias in SET have included adjusting the weighting of student evaluations, contextualizing results with demographic data, and introducing moderated qualitative comment sections. However, these interventions remain inconsistently applied across institutions (Linse, 2017). Scholars such as Peterson and Miles (2024) have called for a complete reconceptualization of how student feedback is collected and used, arguing for models that emphasize developmental feedback over summative judgment.

The emotional toll of biased evaluations has become a prominent theme in contemporary research. Sidwell et al. (2025) found that faculty frequently experienced anxiety, self-doubt, and burnout in response to negative or abusive SET comments. Crimmins et al. (2024) similarly documented ethical concerns among teaching and learning leaders who viewed SET as violating the principle of "first, do no harm." These findings align with Lakeman, Coutts, Hutchinson, Massey, Nasrawi, Fielden, et al. (2022), who linked repeated exposure to hostile student commentary to symptoms of occupational stress disorder. Collectively, these studies illustrate that the consequences of bias are not merely statistical but profoundly human, influencing morale, retention, and institutional climate.

Bias in SET does not merely reflect individual prejudice but is embedded in broader institutional structures. As such, addressing it requires a systemic response that involves faculty training, instrument redesign, and policy reform. Without such efforts, SETs risk reinforcing existing inequities rather than promoting teaching excellence.

In summary, bias in SET is multidimensional, intersectional, and contextually shaped by factors such as anonymity, modality, and language. These forms of bias collectively undermine the validity and fairness of student evaluations, reinforcing systemic inequities within higher education. Although scholars have called for multi-source assessment models that triangulate SET with peer review and teaching portfolios (Hornstein, 2017; Linse, 2017), many institutions continue to rely on student ratings as the primary measure of instructional effectiveness. The persistence of this reliance underscores the importance of examining whether modality-based differences reflect genuine pedagogical variation or embedded bias, a question central to the present study.

Response Rates and Data Quality in SET

Response rates in student evaluations of teaching (SET) are a critical factor affecting the reliability and generalizability of findings. As higher education institutions increasingly transition to online evaluation systems, concerns have emerged regarding declining participation rates and their implications for data quality. Low response rates may introduce non-response bias, reduce the representativeness of results, and undermine the credibility of SET outcomes (Chen & Hoshower, 2021). Student participation in end-of-course evaluations is voluntary, and response patterns may vary across instructional modalities. Chapman and Joines (2017) noted that response rates in online evaluation systems can fluctuate substantially and may influence the representativeness of evaluation data. Lower participation rates, particularly in fully online environments, may limit the extent to which aggregated scores reflect the full population of enrolled students.

Anderson and Soden (2023) analyzed institutional data from over 30,000 course evaluations and found that average response rates in online systems remain 20–30 percentage

points lower than in paper-based formats. Kember et al. (2023) reported similar patterns across Australian universities, noting that voluntary online evaluations rarely exceed 50% participation without incentives. Papadogiannis et al. (2024) observed that post-pandemic students show "evaluation fatigue," completing only mandatory surveys, which diminishes sample representativeness. These findings confirm a long-term trend toward declining participation and underscore the need for methodological safeguards when interpreting SET data.

Studies have consistently found that response rates in online SET systems are significantly lower than in paper-based or in-class evaluations. In a meta-analysis conducted by Avery et al. (2006), response rates for online evaluations were approximately 20–30% lower than those for traditional paper evaluations. More recent research by Liu and Carless (2023) confirms that while digital evaluations offer logistical advantages, they are also associated with increased variability and skewed distributions.

Low response rates increase the likelihood of nonresponse bias, where the views of more satisfied or more dissatisfied students dominate the sample. Research by Fynn (2024) demonstrates that when fewer than half of enrolled students respond, mean SET scores systematically inflate positive extremes. Rivera and Beranek (2023) similarly found that students with higher grades are disproportionately represented among respondents. This imbalance produces an overestimation of instructor effectiveness and reduces the reliability of comparisons across modalities or semesters. In short, lower response rates do not simply yield less data, they yield different data, shaped by self-selection and engagement bias.

The modality of course delivery also impacts response rates. Courses taught entirely online often suffer from particularly low participation in evaluations. A study by Owens and Zhang (2022) found that response rates for asynchronous online courses averaged just 34%,

compared to 62% for face-to-face courses taught by the same instructors. This discrepancy may reflect differences in instructor-student rapport, the perceived relevance of feedback, or the lack of real-time reminders in virtual environments.

Studies highlight that online SET completion is sensitive to the design and timing of survey administration. Anderson and Soden (2023) found that short surveys distributed near the end of the academic term yielded the highest participation, while midterm or extended surveys showed rapid attrition. Sullivan et al. (2024) noted that students' motivation to complete evaluations depends on perceived instructor transparency and follow-up; when students believe feedback will lead to visible course improvements, they are significantly more likely to respond. Institutions that embed SET links directly into learning management systems and send personalized reminders also report modest response increases (Papadogiannis et al., 2024). These factors highlight that response behavior is partially controllable through intentional survey design and communication strategies.

The timing and method of survey administration are important mediators of response behavior. Students are more likely to complete evaluations when they are incentivized, reminded through multiple channels, and assured of confidentiality. Anderson and Soden (2023) showed that personalized reminders from instructors, combined with brief in-class announcements (in hybrid and synchronous online settings), significantly improved response rates without compromising data integrity.

Another emerging issue is the demographic profile of respondents. Several studies suggest that students who are highly satisfied or highly dissatisfied are more likely to complete evaluations, while those with moderate or neutral views are underrepresented (Richardson et al., 2024). This polarization can distort mean scores and reduce the diagnostic value of SET data.

Furthermore, response rates tend to vary by gender, academic standing, and GPA, further complicating the interpretation of aggregate results.

Low response rates also affect the stability of SET scores over time. Fewer responses increase the margin of error and reduce the ability to detect meaningful patterns across course iterations. In response, some institutions have adopted statistical modeling techniques such as generalizability theory and Bayesian estimations to correct for measurement error (Kember et al., 2023). Kember et al. (2023) recommend weighting SET results by course enrollment or student performance distributions to ensure respondent samples better represent the population. Fynn (2024) suggests using multiple imputation to estimate missing data patterns, producing more stable mean estimates. Bayesian hierarchical models have also been employed to correct small-sample variability across instructors or modalities (Rivera & Beranek, 2023). Although these techniques cannot replace robust participation, they enhance fairness and comparability, particularly in institutional datasets used for personnel decisions.

To address these concerns, scholars have proposed strategies to enhance response rates while preserving the authenticity of feedback. These include integrating SETs into course management systems, providing students with evidence that their feedback leads to instructional improvements, and embedding evaluation prompts within final course modules (Berk, 2022). However, these approaches require institutional commitment and faculty engagement to be effective. Declining response rates and data quality challenges in SET present significant limitations for their use in high-stakes decision-making. A more nuanced understanding of participation patterns, combined with targeted interventions and robust statistical adjustments, is necessary to preserve the validity and equity of SET instruments in diverse instructional contexts.

In summary, response rates and data quality remain among the most persistent limitations in SET research. Despite decades of digital innovation, participation continues to decline, raising concerns about representativeness and validity. Studies consistently show that low response rates distort mean scores and undermine comparisons across modalities, course levels, and instructors. Addressing these challenges requires both behavioral interventions (e.g., improved survey design and communication) and statistical corrections (e.g., weighting and imputation). For the present study, awareness of these methodological issues is essential when analyzing SET data across online and face-to-face sections, ensuring that observed differences reflect genuine patterns rather than participation artifacts.

Institutional Use and Policy Implications of SET

Institutions of higher education have long relied on student evaluations of teaching (SET) as a key metric for assessing instructional quality and making personnel decisions such as promotion, tenure, and merit-based salary increases. Despite persistent questions regarding their validity and fairness, SETs remain deeply entrenched in academic culture and policy (Hornstein, 2017; Spooren et al., 2013). This reliance persists even in light of mounting evidence that SET scores are influenced by variables unrelated to teaching effectiveness.

Hornstein (2017) and Pittman (2022) describe SET as a manifestation of the managerial logic that has reshaped higher education over the past two decades. Within this framework, teaching quality becomes quantifiable data, transforming complex pedagogical relationships into performance metrics. Stark and Freishtat (2014) warned that such simplification promotes administrative convenience at the expense of validity and context, often positioning SET scores as proxies for “customer satisfaction.” This managerial approach privileges efficiency and

accountability while marginalizing the formative, developmental purpose for which SET was originally designed.

Universities often present SETs as objective measures of instructional performance, yet numerous scholars caution against their use in high-stakes decision-making. Linse (2017) notes that SET data, while useful for formative feedback, are frequently misinterpreted when applied summatively. Similarly, Stark and Freishtat (2014) argue that SETs lack the psychometric rigor required for personnel evaluation, and their continued use in this context risks unjust consequences for faculty.

Crimmins et al. (2024) applied the principle of "first, do no harm" to the use of SET, arguing that evaluation systems must protect faculty well-being as much as they promote accountability. Their study found that academic leaders frequently recognized SET's psychological and professional risks but felt constrained by institutional reliance on quantitative data. The authors proposed an ethical decision-making framework that balances transparency, proportionality, and support, suggesting that SET should be interpreted within a broader ecosystem of teaching evidence rather than as a stand-alone metric. This perspective aligns with calls from Linse (2017) and Hornstein (2017) for multi-source evaluation systems grounded in fairness and context.

Institutional policies governing the use of SETs vary widely. Some institutions provide guidelines for interpreting results, offer disaggregated data by course type and size, and incorporate peer reviews or teaching portfolios as supplemental measures (Boysen, 2015). Others, however, continue to rely heavily on SET averages without accounting for known biases. This inconsistency exacerbates inequities and undermines the credibility of faculty assessment processes.

A growing body of research links SET practices to faculty stress and burnout, particularly when institutions use evaluations punitively. Lakeman et al. (2023) found that exposure to anonymous and abusive comments led to measurable psychological distress among educators, including symptoms consistent with occupational stress disorder. Sidwell et al. (2025) corroborated these findings, noting that faculty often experience anxiety and self-doubt surrounding evaluation release periods. Both studies argue that institutional policies should explicitly acknowledge the emotional labor of teaching and provide safeguards such as mediated feedback review, debriefing sessions, and counseling support. Without such measures, SET systems risk undermining the very teaching quality they purport to enhance.

Some scholars advocate for eliminating SETs from summative evaluations altogether. Uttl et al. (2017) argue that SETs do not predict student learning outcomes and should be replaced with more direct measures of instructional impact. Other experts propose redesigned instruments that prioritize actionable feedback and emphasize student learning gains rather than satisfaction (Chen & Schunn, 2025). Institutional use of SETs also carries legal and ethical considerations. Cases have emerged where biased evaluations have been used in discriminatory ways, resulting in grievances and litigation (Pittman, 2022). To mitigate such risks, policy frameworks must incorporate safeguards such as bias training for evaluators, contextual analysis of SET data, and appeals processes for faculty. By adopting evidence-based, equity-focused policies, institutions can preserve the value of student feedback while minimizing the risks of misuse and bias. This policy evolution is essential for maintaining both faculty trust and academic integrity in the evaluation of teaching.

Policy scholars increasingly advocate for transparent, participatory approaches to SET reform. Linse (2017) recommends institutional guidelines that clarify SET's formative purpose

and prohibit its exclusive use in high-stakes personnel decisions. Stark and Freishtat (2014) and Hornstein (2017) propose multi-source evaluation frameworks incorporating peer review, self-reflection, and teaching portfolios. More recent initiatives emphasize data transparency, where faculty have access to anonymized aggregate results and can contextualize ratings within course characteristics (Crimmins et al., 2024). Such reforms not only mitigate harm but also rebuild trust between instructors and administrators, reframing SET as a component of continuous improvement rather than an instrument of surveillance.

The institutional use of SET raises profound ethical and governance questions. As universities increasingly rely on student feedback for accountability and quality assurance, the potential for misuse and harm grows correspondingly. Recent scholarship highlights the need for balanced policies that respect faculty dignity, acknowledge emotional labor, and situate SET within a multi-dimensional assessment system. For the current study, understanding these institutional dynamics is critical, as differences in SET scores across modalities may not only reflect pedagogical variation but also the policy environments and evaluative cultures that shape how feedback is interpreted and applied.

Conclusion

This literature review has critically examined student evaluations of teaching (SET) with a particular emphasis on instructional modality, validity, bias, response rates, and institutional use. A broad array of recent and seminal studies were reviewed, including works by Quansah et al. (2024), Chen and Schunn (2025) Uttl et al. (2017), Stark and Freishtat (2014), all of which challenge the psychometric robustness and fairness of SET instruments. These studies serve as pivotal contributions, illuminating how student ratings often fail to reflect teaching quality and are susceptible to contextual and demographic biases.

Over several decades, student evaluations of teaching (SET) have evolved from a formative feedback tool into a summative instrument that influences faculty promotion, tenure, and institutional reputation. The literature consistently shows that while SET offers valuable insights into students' perceptions of teaching effectiveness, its validity and reliability remain contested. Studies reveal that SET scores often reflect factors unrelated to learning outcomes, such as course difficulty, grading leniency, or student expectations (Benton & Li, 2015b; Uttl et al., 2017). Despite technological advances and digital administration, contemporary research continues to raise concerns about psychometric robustness, representativeness, and ethical application.

As higher education adapts to diversified instructional modalities, the relationship between teaching format and SET outcomes has become increasingly important. Comparative studies such as Qin (2025), Aldeman et al. (2025), and Tu and Adkins (2020) indicate that online courses, particularly asynchronous ones, often receive lower evaluations than face-to-face sections taught by the same instructor. However, these differences appear influenced by communication expectations and social perceptions rather than actual instructional quality. Theoretical frameworks such as Expectancy Violation Theory (Burgoon, 1993) and Social Role Theory (Eagly, 1987; Goode, 1960) suggest that students' preconceived norms about interaction, presence, and authority shape their evaluations, especially in technology-mediated settings.

At the same time, SET continues to attract ethical and institutional scrutiny. Recent studies have highlighted bias as a persistent and multifaceted problem, extending beyond gender to include race, language, age, and accent (M. Daskalopoulou, 2024; A. Heffernan, 2022; Lakeman, Coutts, Hutchinson, Massey, Nasrawi, & Fielden, 2022). Intersectional analyses confirm that marginalized faculty often experience compounded disadvantage, particularly in

anonymous online evaluations (A. Daskalopoulou, 2024). Furthermore, research demonstrates that the misuse of SET for high-stakes personnel decisions contributes to stress, emotional exhaustion, and occupational harm (Lakeman et al., 2023; Sidwell et al., 2025). These findings have prompted calls for institutional policies guided by the ethical principle of nonmaleficence, "first, do no harm", and for multi-source evaluation systems incorporating peer review, self-reflection, and qualitative feedback (Crimmins et al., 2024; Linse, 2017).

Despite extensive inquiry, significant gaps remain in the literature.

1. Limited integration of psychometric rigor and contextual analysis. While recent studies employ advanced statistical methods such as generalizability theory and Rasch modeling, few apply these frameworks to institutional data that capture real-world variation across instructional modalities.
2. Insufficient exploration of modality–bias interaction. Research confirms that SET outcomes differ between online and face-to-face environments, but few studies examine how modality amplifies or mitigates bias related to gender, race, or language. Most comparative analyses are limited to pre-pandemic or single-term data, leaving unanswered questions about post-pandemic teaching contexts.
3. Understudied institutional and ethical dimensions. Although recent scholarship acknowledges the emotional and organizational consequences of SET misuse, quantitative evidence connecting these institutional practices to observable rating patterns remains scarce.

These gaps collectively highlight the need for new studies that examine SET through both a psychometric and contextual lens, evaluating not only the statistical differences across modalities but also the social and institutional conditions that shape student perceptions. Addressing these

gaps requires designs that combine quantitative rigor with theoretical grounding in expectancy violation, social role, and institutional theory frameworks.

The current study responds directly to these needs. By analyzing secondary SET data from a state college in Colorado, this research investigates whether significant differences exist in undergraduate SET item scores between online and face-to-face instructional modalities. Through this design, the study aims to determine whether modality-based differences represent genuine pedagogical variation or reflect perceptual and contextual biases embedded within the evaluation process. Findings will contribute to institutional decision-making, ethical policy development, and the broader scholarly understanding of SET as both a measurement tool and a reflection of higher education's evolving instructional landscape.

Chapter 3

Methodology

Introduction

This chapter details the quantitative methodology used to examine differences in student evaluations of teaching (SET) across instructional modalities, online and face-to-face, at a state college in Colorado. Because the study relies on existing institutional data collected as part of routine course evaluation processes, the design is non-experimental and ex post facto. The chapter describes the research design, population and sample, instrumentation, procedures, and analysis plan corresponding to the research question, sub-problems, and hypotheses presented in Chapter 1.

Research Design

This quantitative study uses a non-experimental, ex post facto design to examine differences in student evaluations of teaching (SET) across instructional modalities, i.e., online (asynchronous) or face-to-face, at a state college in Colorado. Using secondary analysis of archival SET data, this study analyzed undergraduate course evaluations collected by the institution over approximately two academic years.

The independent variable is instructional modality, coded as either *online* or *face-to-face*. The dependent variable is the mean Student Evaluation of Teaching (SET) score, derived from the institution's standardized course evaluation instrument. The study compared average SET scores between the two modalities to determine whether statistically significant differences exist. This design is appropriate because it examines naturally occurring data without manipulation or random assignment. The study aligns with prior comparative research examining differences in

SET outcomes across modalities (Benton & Li, 2015a; Macfayden et al., 2023; Rivera & Beranek, 2023; Tu & Adkins, 2020).

Sample

Study participants are represented by a census of 3,611 undergraduate course sections evaluated through the institution's standardized Student Evaluation of Teaching (SET) survey during the study window (approximately two academic years). The data are de-identified and aggregated at the course-section level. Each record reflects the mean scores of student responses within that course section for each survey question. No individual student or instructor identifiers were included.

Inclusion Criteria

- Undergraduate courses at the college.
- Course sections formally designated by the registrar as online (asynchronous) or face-to-face.
- Courses evaluated using the same institutional SET instrument across both modalities.
- Sections with at least five valid responses to ensure representativeness of results.
- Courses offered during the most recent two academic years for which complete data are available.

Exclusion Criteria

- Hybrid or HyFlex courses (partially online).
- Noncredit, dual-enrollment, or independent study courses.
- Sections with fewer than five responses or missing key variable data (i.e., modality code or item means).
- Sections evaluated with a different or experimental instrument version.

Recruitment (Data Access)

Because this study uses existing institutional data, no recruitment of human participants occurred. The researcher obtained access through the Office of Institutional Effectiveness (OIE) following written approval and IRB exemption. The dataset included only de-identified, section-level data stripped of instructor and student identifiers.

Instrumentation

The college uses a standardized Student Evaluation of Teaching (SET) survey to collect feedback on instructional quality at the end of each term. The survey is administered electronically through the college's learning management or survey platform and is consistent across all modalities. The items measure student perceptions of course organization, instructor preparedness, engagement, and learning support. See Appendix A for the items on the SET.

The instrument consists of 12 Likert-scale items rated on a 4-point scale:
1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree.
The minimum score for the SET is 12, while the maximum score is 48. Thus, the mean scores ranged from 1 to 4, with higher scores indicative of a more positive rating.
Qualitative open-ended questions appearing at the end of the survey are excluded from analysis.

Reliability and Validity

Prior analyses at other institutions indicate strong internal consistency for the instrument (Cronbach's α typically $\geq .85$). In addition, the study recomputed Cronbach's α for the current dataset to document internal consistency. This is consistent with psychometric expectations for SET instruments reported in the literature (Kember et al., 2023; Marsh, 1987; Spooren et al., 2013; Tavakol & Dennick, 2011). While SET captures student perceptions (not direct learning

outcomes), it is widely used as a standardized indicator of perceived teaching quality across modalities (Benton & Li, 2015a; Rivera & Beranek, 2023).

Demographics

Each SET record includes section-level metadata such as:

- Course term and year.
- Course modality (online or face-to-face).
- Class enrollment count and number of respondents.

No identifying demographic variables (i.e., student age, or program) were used in this study.

Procedures

The research followed these sequential steps:

1. IRB and Institutional Permissions
    - Obtain IRB exempt review approval from Marywood University.
    - Request formal approval from the college's Institutional Review Board or Exempt Committee and Institutional Effectiveness office for access to de-identified data.
2. Data Request and Acquisition
    - Submit a written request specifying data elements: modality, term, section-level mean scores for each of the 12 Likert items, total enrollment, and number of respondents.
    - Data was extracted and de-identified by OIE staff prior to transfer.
3. Data Storage and Security
    - Data was stored on an encrypted institutional drive accessible only to the researcher.
    - No data was transmitted via email or stored on personal or cloud devices.

4. Data Cleaning and Preparation
    - Prior to conducting statistical analyses, the dataset was reviewed and prepared to ensure compatibility with SPSS and to support valid and reliable quantitative analysis. Data preparation focused on variable formatting, recoding, labeling, and verification of analytic assumptions.
5. Variable Coding and Formatting
    - All variables intended for statistical analysis were converted to numeric formats, as SPSS requires numeric values for inferential procedures such as correlations and t-tests. Categorical variables originally represented with text labels were recoded using numeric values, with value labels assigned to preserve interpretability. For example, instructional modality was coded as 0 = online and 1 = face-to-face, instructor gender was coded using numeric categories consistent with the institutional dataset, and course level was coded as 1 = 1000-level and 2 = 2000-level. This approach allowed categorical variables to be used as grouping variables in independent samples t-tests.
    - Descriptive variable names and labels were reviewed and standardized to improve clarity in SPSS output. Variable labels were applied so that tables and output clearly reflected the meaning of each variable rather than abbreviated field names. This ensured interpretability of results during analysis and reporting.
6. Filtering and Subgroup Analyses
    - To support comparisons across instructional modalities and other grouping variables, the dataset was filtered as needed using SPSS case selection procedures. Analyses were conducted separately for online and face-to-face courses by

applying conditional filters prior to running correlations and *t*-tests. This approach ensured that subgroup analyses reflected only the relevant cases and prevented cross-contamination of results between instructional modalities.

- Filters were systematically applied and removed between analyses to allow repeated testing across different subgroups while maintaining a single, unified dataset.

7. Creation of Derived Variables
    - Additional variables were computed to support specific research questions. Section-level survey response rate was calculated as the percentage of enrolled students who completed the SET survey. Because response counts varied by item, the maximum response count across SET items was used to represent the total number of student responses per section. This value was divided by section enrollment and multiplied by 100 to yield a percentage-based response rate variable suitable for comparison across instructional modalities.
    - Derived variables were verified through manual checks to confirm accurate calculation and consistency with raw data values.
8. Data Verification and Assumption Review
    - Throughout the data preparation process, frequency distributions, descriptive statistics, and summary tables were reviewed to identify potential anomalies, missing data issues, or out-of-range values. No data transformations were required beyond recoding and variable creation, as the analyses selected were robust to minor deviations from normality given the large sample size.

- All cleaned and derived variables were saved within the primary SPSS data file to preserve reproducibility and to ensure consistency across analyses.

9. Data Analysis and Retention
    - Data was analyzed using IBM SPSS Statistics (Version 30).
    - Following analysis, all files will be securely deleted within one year of dissertation defense.

Data Analysis

Data was analyzed at the course-section level. The significance level for all tests is set at $\alpha = .05$ (two-tailed).

Sub-Problem 1, *What are the SET scores in online classes*, was analyzed using a frequency distribution and other descriptive statistics.

Sub-Problem 2, *What are the SET scores in face-to-face classes*, was analyzed using a frequency distribution and other descriptive statistics.

Sub-Problem 3, *What are the differences in SET scores across modalities*, was analyzed using an independent samples t-test.

Supplemental Analysis

Descriptive comparisons were conducted by course discipline, course level, and gender of the instructor.

Conclusion

This chapter outlines the quantitative design, sample, instrumentation, data procedures, and analysis plan for this study. Using archival course evaluation data from a Colorado college, the research examined whether instructional modality influences student evaluations of teaching. The findings will contribute to understanding how SET outcomes differ across learning

environments and inform institutional practices for fair and valid evaluation of teaching effectiveness.

Chapter 4

Results

Introduction

This chapter presents the results of the quantitative analyses conducted to examine differences in undergraduate Student Evaluations of Teaching (SET) across instructional modalities, online and face-to-face. The analyses address the primary research question and associated subproblems outlined in Chapters 1 and 3. Data were analyzed at the course-section level using IBM SPSS Statistics version 30. Statistical significance was evaluated at the $\alpha = .05$ level using two-tailed tests.

Description of the Sample

The final dataset included 3,611 undergraduate course sections that met all inclusion criteria. Of these sections:

- 1,601 were delivered online
- 2,010 were delivered face-to-face

All sections used the institution's standardized 12-item SET instrument and included a minimum of five student responses. All data were aggregated at the section level and de-identified prior to analysis.

Description of the Instrument

The instrument used in this study was a 12-item Student Evaluation of Teaching (SET) survey (Appendix A) administered institution-wide. Each item was measured using a 4-point Likert-type scale, with higher values indicating more favorable student perceptions of instructional effectiveness. For purposes of analysis, a composite SET score was calculated by

averaging the 12 item means for each course section. This composite score served as the primary dependent variable in subsequent analyses examining differences across instructional modalities.

Reliability of the SET Instrument

Internal consistency reliability was examined separately for each instructional modality. The 12-item SET instrument demonstrated excellent internal consistency for online courses (Cronbach's $\alpha = .98$) and for face-to-face courses (Cronbach's $\alpha = .977$). These values exceed the commonly accepted .70 threshold for acceptable reliability and support the use of composite SET scores in subsequent analyses.

Sub-Problem 1, What are the SET scores in online classes, was analyzed using a frequency distribution and descriptive statistics.

The mean composite SET scores for the online courses was 3.57 (SD = 0.26), while the median was also 3.57, range = 3.52-3.68. Table 1 presents the item-level means and standard deviations for the 12 SET items in online courses.

Questions	N	Mean	SD
Q1 My instructor helped me engage with the course material	1601	3.52	0.31
Q2 My instructor was well prepared for each class.	1601	3.61	0.30
Q3 My instructor communicated ideas and concepts clearly.	1601	3.52	0.33
Q4 My instructor organized this course well.	1601	3.55	0.32
Q5 My instructor demonstrated knowledge of the subject.	1601	3.68	0.24
Q6 My instructor facilitated productive class discussions and activities.	1601	3.57	0.28
Q7 The feedback from my instructor assisted my learning.	1601	3.53	0.32
Q8 My instructor was regularly available for consultation (e.g., responsive to email).	1601	3.56	0.32
Q9 My instructor encouraged an atmosphere of mutual respect.	1601	3.66	0.22
Q10 My instructor helped me feel that I belonged in this course.	1601	3.58	0.29
Q11 Assignments helped me to learn the course content and skills.	1601	3.58	0.26
Q12 Course materials were useful in my learning and relevant to this course.	1601	3.60	0.25
Composite Score	1601	3.57	0.26

Table 1: Item-level means and std. deviations for 12 SET items in online courses

The highest mean scores were observed for Q5: My instructor demonstrated knowledge of the subject (M = 3.68, SD = 0.24), Q9: My instructor encouraged an atmosphere of mutual

respect (M = 3.66, SD = 0.22), Q2: My instructor was well prepared for each class (M = 3.61, SD = 0.30), and Q12: Course materials were useful in my learning and relevant to this course (M = 3.60, SD = 0.25).

The lowest mean scores were observed for Q1: My instructor helped me engage with the course material (M = 3.52, SD = 0.31), Q3: My instructor communicated ideas and concepts clearly (M = 3.52, SD = 0.33), Q7: The feedback from my instructor assisted my learning (M = 3.53, SD = 0.32), and Q4: My instructor organized this course well (M = 3.55, SD = 0.32).

Overall, item-level means, indicated consistently high ratings across all online instructional items.

Sub-Problem 2, What are the SET scores in face-to-face classes, was analyzed using a frequency distribution and other descriptive statistics.

The mean composite SET scores for the face-to-face courses was 3.63 (SD = 0.27), while the median was also 3.63, range = 3.58-3.80. Table 2 presents the item-level means and standard deviations for the 12 SET items in face-to-face courses.

Questions	N	Mean	SD
Q1 My instructor helped me engage with the course material	2010	3.62	0.33
Q2 My instructor was well prepared for each class.	2010	3.68	0.30
Q3 My instructor communicated ideas and concepts clearly.	2010	3.59	0.36
Q4 My instructor organized this course well.	2010	3.59	0.35
Q5 My instructor demonstrated knowledge of the subject.	2010	3.80	0.21
Q6 My instructor facilitated productive class discussions and activities.	2010	3.63	0.35
Q7 The feedback from my instructor assisted my learning.	2010	3.59	0.34
Q8 My instructor was regularly available for consultation (e.g., responsive to email).	2010	3.63	0.29
Q9 My instructor encouraged an atmosphere of mutual respect.	2010	3.72	0.25
Q10 My instructor helped me feel that I belonged in this course.	2010	3.65	0.31
Q11 Assignments helped me to learn the course content and skills.	2010	3.58	0.32
Q12 Course materials were useful in my learning and relevant to this course.	2010	3.60	0.30
Composite Score	2010	3.63	0.27

Table 2: Item-level means and std. deviations for 12 SET items in face-to-face courses

The mean composite SET score for face-to-face courses was 3.63 (SD = 0.27). The highest mean scores were observed for Q5: My instructor demonstrated knowledge of the subject (M = 3.80, SD = 0.21), Q9: My instructor encouraged an atmosphere of mutual respect (M = 3.72, SD = 0.25), Q2: My instructor was well prepared for each class (M = 3.68, SD = 0.30), and Q10: My instructor helped me feel that I belonged in this course (M = 3.65, SD = 0.31).

The lowest mean scores were observed for Q11: Assignments helped me to learn the course content and skills (M = 3.58, SD = 0.32), Q3: My instructor communicated ideas and concepts clearly (M = 3.59, SD = 0.36), Q4: My instructor organized this course well (M = 3.59, SD = 0.35), and Q7: The feedback from my instructor assisted my learning (M = 3.59, SD = 0.34).

Overall, item-level means, indicated consistently high ratings across all face-to-face instructional items, with slightly higher overall ratings compared to the online modality.

Sub-Problem 3, What are the differences in SET scores across modalities, was analyzed using an independent samples t-test.

Prior to conducting independent samples t-tests, Levene's test for equality of variances was examined to determine whether equal variance assumptions were met. In cases where equal variances were not assumed, adjusted degrees of freedom were reported. Given the large sample size in both instructional modalities, minor deviations from normality were not considered problematic.

An independent samples t-test revealed a statistically significant difference in composite SET scores between online and face-to-face modalities, $t(3609) = -6.88$, $p < .001$. However, the effect size was small ($d = 0.23$), indicating that the magnitude of the difference between instructional modalities was modest. Face-to-face courses received higher composite SET scores

than online courses. Although statistically significant, the effect size was small, indicating modest practical differences between instructional modalities. The null hypothesis, there would be no significant differences in SET scores across instructional modalities was rejected. These results indicate that statistically significant differences exist in SET scores between online and face-to-face instructional modalities.

Supplemental Analysis

Intercorrelations Among SET Items by Modality

Pearson correlation analyses were conducted separately for online and face-to-face courses to examine the degree of intercorrelation among the 12 Student Evaluation of Teaching (SET) items.

Online Courses

	Q2 Mean	Q3 Mean	Q4 Mean	Q5 Mean	Q6 Mean	Q7 Mean	Q8 Mean	Q9 Mean	Q10 Mean	Q11 Mean	Q12 Mean
Q1 My instructor helped me engage with the course material	.845**	.884**	.838**	.793**	.844**	.851**	.788**	.798**	.853**	.795**	.772**
Q2 My instructor was well prepared for each class.		.861**	.873**	.787**	.793**	.797**	.795**	.757**	.797**	.747**	.733**
Q3 My instructor communicated ideas and concepts clearly.			.867**	.795**	.817**	.855**	.786**	.764**	.828**	.784**	.766**
Q4 My instructor organized this course well.				.753**	.808**	.809**	.763**	.735**	.788**	.784**	.778**
Q5 My instructor demonstrated knowledge of the subject.					.788**	.772**	.726**	.775**	.759**	.729**	.716**
Q6 My instructor facilitated productive class discussions and activities.						.820**	.732**	.791**	.816**	.799**	.764**
Q7 The feedback from my instructor assisted my learning.							.790**	.788**	.849**	.781**	.767**
Q8 My instructor was regularly available for consultation (e.g., responsive to email)								.764**	.775**	.684**	.678**
Q9 My instructor encouraged an atmosphere of mutual respect.									.845**	.744**	.729**
Q10 My instructor helped me feel that I belonged in this course.										.810**	.776**
Q11 Assignments helped me to learn the course content and skills.											.864**
Q12 Course materials were useful in my learning and relevant to this course.											

**All of the correlations are significant at the .001 level.

Table 3: Inter-correlation among SET items in online classes

For online courses, inter-item correlations among the SET items were uniformly strong. Table 3 shows correlation coefficients ranged from r = .77 to r = .88, with all correlations statistically significant at $p < .001$. Correlations between individual SET items and the composite SET score were also strong, ranging from r = .85 to r = .95, all significant at $p < .001$. These results indicate a high degree of internal consistency among SET items within the online instructional modality.

Face-to-Face Courses

Similarly, for face-to-face courses, inter-item correlations demonstrated strong positive relationships across all SET items, as demonstrated in Table 4. For face-to-face courses, inter-item correlations ranged from r = .71 to r = .89, with all correlations statistically significant at p < .001.These findings indicate substantial internal consistency among SET items within face-to-face courses.

	Q2 Mean	Q3 Mean	Q4 Mean	Q5 Mean	Q6 Mean	Q7 Mean	Q8 Mean	Q9 Mean	Q10 Mean	Q11 Mean	Q12 Mean
Q1 My instructor helped me engage with the course material	.772**	.873**	.820**	.746**	.889**	.864**	.707**	.787**	.863**	.810**	.781**
Q2 My instructor was well prepared for each class.		.786**	.832**	.723**	.776**	.714**	.654**	.702**	.729**	.695**	.682**
Q3 My instructor communicated ideas and concepts clearly.			.842**	.725**	.862**	.833**	.675**	.779**	.833**	.795**	.751**
Q4 My instructor organized this course well.				.690**	.823**	.796**	.670**	.752**	.796**	.792**	.769**
Q5 My instructor demonstrated knowledge of the subject.					.759**	.713**	.629**	.725**	.717**	.664**	.646**
Q6 My instructor facilitated productive class discussions and activities.						.863**	.691**	.800**	.840**	.833**	.803**
Q7 The feedback from my instructor assisted my learning.							.731**	.808**	.859**	.798**	.761**
Q8 My instructor was regularly available for consultation (e.g., responsive to email)								.706**	.735**	.656**	.639**
Q9 My instructor encouraged an atmosphere of mutual respect.									.861**	.763**	.755**
Q10 My instructor helped me feel that I belonged in this course.										.849**	.816**
Q11 Assignments helped me to learn the course content and skills.											.866**
Q12 Course materials were useful in my learning and relevant to this course.											

**All of the correlations are significant at the .001 level.

Table 4:Inter-correlation among SET items in face-to-face classes

Across both instructional modalities, SET items exhibited consistently strong intercorrelations, suggesting that students tended to respond similarly across items regardless of item-specific content. The magnitude and consistency of these correlations support the use of a composite SET score in subsequent analyses.

Relationship Between SET Scores and Academic Outcomes

Pearson correlation analyses were conducted to examine relationships between Student Evaluation of Teaching (SET) scores and academic outcomes, including section-level grade point average (GPA) and the percentage of A and B grades awarded. Analyses were conducted separately for online and face-to-face courses using the composite SET score.

Online Courses

		GPA	%AB
Composite Score	Pearson Correlation	0.25**	0.25**
	N	1601	1601
GPA	Pearson Correlation		0.96**
	N		1601

**correlation is significant at the 0.01 level (2-tailed)

Table 5: Composite SET score in online classes

For online courses, Table 5 shows the composite SET score demonstrated a positive relationship with section-level GPA ($r = .25$, $p < .001$) and with the percentage of A and B grades awarded ($r = .25$, $p < .001$). A very strong positive correlation was observed between GPA and the percentage of A and B grades ($r = .96$, $p < .001$).

Face-to-Face Courses

		GPA	%AB
Composite Score	Pearson Correlation	0.28**	0.27**
	N	2010	2010
GPA	Pearson Correlation		0.96**
	N		2010

**correlation is significant at the 0.01 level (2-tailed)

Table 6: Composite SET score in face-to-face classes

For face-to-face courses, the composite SET score was positively correlated with section-level GPA ($r = .28$, $p < .001$) and with the percentage of A and B grades ($r = .21$, $p < .001$). As with online courses, GPA and the percentage of A and B grades were strongly correlated ($r = .96$, $p < .001$).

Across both instructional modalities, composite SET scores demonstrated statistically significant positive relationships with academic performance outcomes. In both online and face-to-face courses, higher SET scores were associated with higher section-level GPA and a greater proportion of A and B grades awarded. The relationship between GPA and grade distribution was strong and consistent across modalities.

Differences in SET Scores by Instructional Modality

Independent samples *t*-tests were conducted to compare online and face-to-face course sections on each of the 12 Student Evaluation of Teaching (SET) items and on the composite SET score. The null hypothesis for each questions was, “There is no difference between the mean scores for question X for online vs. face-to-face classes (where X represents each of the 12 questions and the composite). The alternate hypothesis for each question was “There is a difference between the mean scores for question X for online vs. face-to-face classes (where X represents each of the questions).

Individual SET Items

Statistically significant differences favoring face-to-face courses were observed for Items 1 through 10, with all significant comparisons demonstrating *p* values less than .001. Representative results include:

	Online Mean	F2F Mean	Sig. Value	Cohen's d
Q1 My instructor helped me engage with the course material	3.51	3.62	0.000	0.333
Q2 My instructor was well prepared for each class.	3.60	3.68	0.000	0.238
Q3 My instructor communicated ideas and concepts clearly.	3.52	3.59	0.000	0.205
Q4 My instructor organized this course well.	3.55	3.59	0.000	0.119
Q5 My instructor demonstrated knowledge of the subject.	3.68	3.81	0.000	0.585
Q6 My instructor facilitated productive class discussions and activities.	3.57	3.63	0.000	0.172
Q7 The feedback from my instructor assisted my learning.	3.53	3.59	0.000	0.184
Q8 My instructor was regularly available for consultation (e.g., responsive to email).	3.56	3.63	0.000	0.236
Q9 My instructor encouraged an atmosphere of mutual respect.	3.66	3.72	0.000	0.267
Q10 My instructor helped me feel that I belonged in this course.	3.58	3.65	0.000	0.220
Q11 Assignments helped me to learn the course content and skills.	3.58	3.58	0.637	0.015
Q12 Course materials were useful in my learning and relevant to this course.	3.59	3.60	0.661	0.014
Composite Score	3.57	3.63	0.000	0.231

Table 7: Differences in SET scores by modality

Across the statistically significant items, effect sizes ranged from small to moderate, with Cohen’s *d* values ranging from 0.23 to 0.53. Item-level analyses further supported rejection of

the null hypothesis, as statistically significant differences were observed across multiple SET items.

Overall, face-to-face courses received higher SET scores than online courses on the composite measure and on most individual items. Differences were statistically significant for the majority of SET items, though the magnitude of these differences varied across items.

Instructor Gender and SET Scores

Independent samples *t*-tests were conducted to examine differences in perceived subject matter knowledge, as measured by Item 5 of the Student Evaluation of Teaching (SET) instrument, by instructor gender within each instructional modality. Instructor gender was coded as a binary variable in the institutional dataset; therefore, results reflect differences only within the available gender categories.

For online classes, there was no statistically significant difference in Q5 ratings between male instructors (M = 3.69, SD = 0.25, n = 508) and female instructors (M = 3.67, SD = 0.23, n = 1093), t(1599) = 1.13, p = .258.

For face-to-face sections, male instructors (M = 3.81, SD = 0.20, n = 925) and female instructors (M = 3.80, SD = 0.21, n = 1085), t(1992.67) = 1.56, p = .12, Cohen's d = 0.69. The observed mean difference of .01 was minimal and the effect size was trivial.

The independent samples *t*-tests indicated no statistically significant differences in perceived instructor subject matter knowledge by instructor gender in either online or face-to-face courses. Although mean ratings were slightly higher for male instructors in both instructional modalities, these differences were small in magnitude and did not reach statistical significance. Overall faculty gender does not appear to be a significant factor in regards to the composite SET scores received across the combined instructional modalities.

Course Level Differences in SET Scores

	Online Mean	F2F Mean
Q1 My instructor helped me engage with the course material		
1000 Level	3.50	3.61
2000 Level	3.51	3.65
Sig. Value	0.38	0.09**
Q2 My instructor was well prepared for each class.		
1000 Level	3.60	3.67
2000 Level	3.59	3.70
Sig. Value	0.62	0.03
Q3 My instructor communicated ideas and concepts clearly.		
1000 Level	3.51	3.58
2000 Level	3.51	3.62
Sig. Value	0.90	0.08
Q4 My instructor organized this course well.		
1000 Level	3.54	3.58
2000 Level	3.51	3.60
Sig. Value	0.10	0.27
Q5 My instructor demonstrated knowledge of the subject.		
1000 Level	3.66	3.80
2000 Level	3.68	3.83
Sig. Value	0.14	0.01**
Q6 My instructor facilitated productive class discussions and activities.		
1000 Level	3.56	3.63
2000 Level	3.58	3.62
Sig. Value	0.34	0.95
Q7 The feedback from my instructor assisted my learning.		
1000 Level	3.51	3.58
2000 Level	3.53	3.61
Sig. Value	0.39	0.29
Q8 My instructor was regularly available for consultation (e.g., responsive to email).		
1000 Level	3.55	3.62
2000 Level	3.57	3.69
Sig. Value	0.29	0.00**
Q9 My instructor encouraged an atmosphere of mutual respect.		
1000 Level	3.65	3.72
2000 Level	3.66	3.74
Sig. Value	0.46	0.23
Q10 My instructor helped me feel that I belonged in this course.		
1000 Level	3.57	3.65
2000 Level	3.58	3.65
Sig. Value	0.90	0.85
Q11 Assignments helped me to learn the course content and skills.		
1000 Level	3.57	3.57
2000 Level	3.58	3.59
Sig. Value	0.50	0.31
Q12 Course materials were useful in my learning and relevant to this course.		
1000 Level	3.59	3.59
2000 Level	3.59	3.61
Sig. Value	0.92	0.17

*Table 8: t-test lower level vs. upper level by modality; **correlation is significant at the 0.01 level (2-tailed)*

Independent samples t-tests were conducted to compare SET scores between lower-division (1000-level) and upper-division (2000-level) courses within each instructional modality.

Independent samples t-tests were conducted to examine differences between 1000-level and 2000-level courses within each modality. No statistically significant differences were observed within the online modality across any of the 12 items. However, three statistically significant differences emerged within the face-to-face modality.

Statistically significant differences favoring upper-division courses were observed for:

- Item 2, $p < .05$, Cohen's $d = 0.03$
- Item 5, $p < .05$, Cohen's $d = 0.09$
- Item 8, $p < .05$, Cohen's $d = 0.60$
- For question 2 in face-to-face sections, 2000-level courses (M = 3.70) received slightly higher ratings than 1000-level courses (M = 3.67), p = .03. Although statistically significant, the mean difference of .03 is small. This finding suggests that students in upper-level face-to-face courses perceived instructors as slightly more prepared compared to those in lower-level face-to-face courses.
- For question 5 in face-to-face sections, 2000-level courses (M = 3.83) were rated higher than 1000-level courses (M = 3.80), p = .01. The mean difference was .03. Although the statistical test indicated significance, the magnitude of the difference was minimal. Ratings were high for both levels, indicating that students perceived instructors as knowledgeable regardless of course level, with only a slight increase in upper-level face-to-face courses.
- For question 8 in face-to-face sections, 2000-level courses (M = 3.69) were rated higher than 1000-level courses (M = 3.62), p < .01. The mean difference was .07. While this was

the largest of the three significant differences, the absolute difference remains small on a 4-point scale. This result indicates that students in upper-level face-to-face courses perceived slightly greater instructor availability compared to students in lower-level face-to-face courses.

Course level differences in SET scores were limited to specific items in face-to-face courses, while no course level differences were observed in online courses.

Course Section Size and SET Scores

Pearson correlation analyses were conducted to examine relationships between section enrollment and Student Evaluation of Teaching (SET) scores. Although several correlations reached statistical significance, all observed relationships were small in magnitude.

Online Courses

For online courses, section enrollment demonstrated weak negative correlations with two SET items:

- Item 5: $r = -.07$, $p = .003$
- Item 7: $r = -.06$, $p = .02$

No statistically significant relationship was observed between section enrollment and the composite SET score. All other correlations between enrollment and individual SET items were non-significant.

Face-to-Face Courses

For face-to-face courses, section enrollment was negatively correlated with all individual SET items as well as the composite SET score. Correlation coefficients ranged from $r = -.06$ to $r = -.10$, with all correlations statistically significant ($p < .001$). The correlation between section enrollment and the composite SET score was $r = -.10$, $p < .001$.

Despite statistical significance, the magnitude of these relationships was small, indicating minimal practical association between class size and SET scores.

Overall, larger section enrollment was associated with slightly lower SET scores, particularly in face-to-face courses; however, the strength of these relationships was weak, suggesting limited practical impact.

Student Response Rates by Instructional Modality

Modality		N	Mean
PctRespond	Online	1601	45.98
	Face-to-Face	2010	52.95

Table 9: Percent response rate by modality

A response rate variable was calculated as the percentage of enrolled students who completed the Student Evaluation of Teaching (SET) survey. Independent samples *t*-tests were conducted to compare response rates between instructional modalities.

Results indicated a statistically significant difference in response rates between online and face-to-face courses:

This difference was statistically significant, $t(3609) = -12.26$, $p < .001$, with a moderate effect size (Cohen's $d = 0.42$). This difference aligns with prior findings that face-to-face courses tend to yield higher evaluation participation rates. This moderate effect size suggests that instructional modality meaningfully influences student participation in course evaluations.

Summary of Results

Results indicated statistically significant differences in Student Evaluation of Teaching (SET) scores across instructional modalities, with face-to-face courses receiving higher ratings on most items and on the composite score. SET items were highly intercorrelated in both instructional modalities. Composite SET scores demonstrated small but statistically significant

relationships with grade point average and grade distributions. Course level differences were limited to specific SET items in face-to-face courses. Section size was negatively associated with SET scores in face-to-face courses, while relationships in online courses were minimal. Response rates were significantly higher in face-to-face courses.

Chapter 5

Discussion

Introduction

The purpose of this study was to examine differences in undergraduate student evaluations of teaching (SET) across instructional modalities, specifically online and face-to-face sections, at a community college in Colorado. Chapter 4 presented descriptive statistics and inferential analyses comparing composite SET scores across modalities. This chapter situates those findings within the broader SET literature, examines the extent to which the results support or contradict prior research, and discusses implications for institutional policy and future research.

Student evaluations of teaching remain one of the most widely used mechanisms for assessing instructional quality in higher education. Despite their prevalence, extensive scholarship has raised concerns regarding sources of measurement error, rater bias, and the appropriateness of high-stakes use without additional validity evidence (Quansah et al., 2024). In addition to validity concerns, scholars have raised questions regarding reliability interpretation (Morely, 2014), grade-related influences on evaluation outcomes (Stroebe, 2016), and modality-based differences in student perception (Rovai et al., 2006).

Three principal findings emerged from this study. First, both online and face-to-face sections received generally high SET scores, with limited variability across sections. Second, face-to-face sections demonstrated a statistically significant but small increase in mean SET scores compared to online sections. Third, the SET instrument demonstrated extremely high internal consistency reliability within this dataset. These findings require careful interpretation, particularly in light of the broader psychometric and critical literature on student ratings.

Interpreting the Small Effect Size

The large sample size (N = 3,611 sections) increases statistical power, making it possible to detect even small differences. Therefore, the practical meaning of the effect size is more important than the p-value. A Cohen's d of 0.23 indicates substantial overlap between score distributions and minimal separation between modality-based patterns, suggesting limited practical differentiation in perceived instructional effectiveness.

In institutional contexts, differences of .05 to .10 on a 4-point scale are often treated as meaningful in tenure and promotion reviews. However, the present findings suggest that such differences may reflect structural modality effects rather than meaningful instructional disparities. Thus, while modality effects are real, their magnitude does not justify strong evaluative distinctions between online and face-to-face faculty.

Expectancy Violation Theory and Modality

Expectancy Violation Theory (EVT) provides a useful interpretive framework for understanding potential differences in student evaluations across instructional modalities. Students often enter higher education with established expectations regarding what constitutes effective teaching, expectations that have historically been shaped by traditional face-to-face educational experiences. These experiences typically emphasize visible instructor presence, immediate verbal and nonverbal feedback, physical classroom interaction, and spontaneous discussion. Research comparing online and face-to-face instruction suggests that differences in interaction patterns and perceived instructor presence can influence student perceptions of course quality (Rovai et al., 2006; Young & Duncan, 2014). In online environments, particularly asynchronous formats, communication norms shift toward text-based feedback, delayed interaction, and reduced nonverbal cues. Such structural differences may alter students'

perceptions of immediacy and relational engagement, even when instructional design and learning outcomes are comparable. Studies examining modality-based rating differences have noted that online instruction may be evaluated differently due to variations in communication channels and student expectations rather than substantive differences in instructional competence (Aldeman et al., 2025; Arias et al., 2018).

In the present study, the observed effect size between modalities was small, suggesting that any expectancy-related differences were modest rather than pronounced. This finding is consistent with more recent research indicating that students have become increasingly familiar with digital learning environments and may adapt their evaluative standards accordingly (Qin, 2025). Although subtle preference patterns may persist, the limited magnitude of the modality difference suggests that expectancy violations, if present, are not severe within this institutional context.

Social Role Theory and Perception

Social Role Theory suggests that evaluative judgments are shaped by socially constructed expectations regarding authority, warmth, and competence, which may influence how instructors are perceived in classroom settings. In technology-mediated environments, such as online courses, the transmission and interpretation of these social signals may differ from face-to-face contexts, potentially altering student perceptions. Although the present dataset did not reveal statistically significant gender differences in composite SET scores, this finding does not negate the broader body of literature documenting gender bias in student evaluations.

Empirical research has consistently demonstrated that student ratings may vary as a function of instructor gender, even when instructional performance is held constant (Boring, 2017; MacNell et al., 2015). The absence of statistically significant gender effects in this study

may instead reflect methodological or contextual factors, including aggregation at the section level rather than analysis of individual-level student responses, institutional culture moderating potential bias effects, limitations inherent in the measurement instrument, or the composition of the sample itself. As research examining intersectional and contextual influences suggests, bias in student evaluations is not uniform across institutions or settings (A. Daskalopoulou, 2024). Accordingly, the present findings underscore the importance of contextual interpretation and caution against generalizing gender-related conclusions beyond the specific institutional environment examined in this study.

Modality Differences

The finding that face-to-face sections received slightly higher mean SET scores than online sections is consistent with prior research indicating that modality can influence student ratings. Some comparative studies have found that traditional in-person courses receive higher overall evaluations than online courses, even when course content is equivalent. However, the literature does not demonstrate uniform agreement regarding the magnitude or meaning of such differences.

Rovai et al. (2006) reported that students evaluating online courses may demonstrate greater extremity in rating behavior, suggesting that response context influences evaluation patterns. Similarly, comparative analyses of modality have shown that differences in perceived instructor presence and communication clarity can shape evaluation outcomes independent of learning performance. These findings are consistent with the small but statistically significant difference observed in the present study.

Although a statistically detectable difference exists, the practical separation between modalities remains limited. This aligns with prior research suggesting that instructional design,

perceived instructor presence, and communication patterns may influence student ratings more directly than delivery format alone. Rather, student expectations, communication structures, perceived instructor presence, and workload design may mediate rating outcomes.

Online instruction often shifts the visibility of teaching behaviors. Elements such as clarity of written communication, timeliness of feedback, and structured organization become more salient in virtual environments. If students interpret immediacy and spontaneous interaction in face-to-face settings as indicators of instructor effectiveness, this may advantage traditional sections in rating processes even when instructional quality is comparable.

Although response rates were not analyzed at the individual student level in this study, the potential for differential participation across modalities should be considered when interpreting findings. Variability in response patterns may introduce limitations in generalizability, particularly if students who choose to respond differ systematically from those who do not. Accordingly, conclusions drawn from aggregated SET scores should be interpreted within the context of voluntary participation and institutional administration practices.

Thus, the present findings suggest that statistically significant modality differences may reflect structural characteristics of course delivery rather than substantive differences in instructional effectiveness.

Reliability and Internal Consistency

The SET instrument in this study demonstrated extremely high internal consistency reliability. High coefficient alpha values are often interpreted as evidence that items measure a cohesive construct. However, psychometric literature cautions that very high alpha values may reflect item redundancy or a narrow construct repeatedly assessed rather than comprehensive measurement, internal consistency estimates increase as the number of highly correlated items

increases (Tavakol & Dennick, 2011). Consequently, high internal consistency reliability does not, by itself, establish construct validity or justify high-stakes interpretive use (Tavakol & Dennick, 2011; Zhao & Gallant, 2012).

Morely (2014) argued that reliance on internal consistency estimates alone may overstate the dependability of student ratings for evaluative purposes, particularly when inter-rater agreement is not separately assessed. This concern is reinforced by Zhao and Gallant (2012) who emphasized that reliability coefficients do not eliminate measurement error arising from contextual and perceptual influences. Accordingly, the high alpha coefficient observed in the present study should be interpreted as evidence of internal cohesion rather than proof of comprehensive construct validity.

Halo Effects, Leniency, and Measurement Error

The clustering of SET scores near the upper end of the scale mirrors patterns documented in prior research. Scholars have identified, as documented in systematic reviews of SET validity evidence, halo effects and rater severity differences as common sources of measurement distortion in student evaluations (Quansah et al., 2024). When students form an overall impression of an instructor and rate all items similarly, dimensional distinctions collapse into a generalized evaluation. Additionally, research has documented positive associations between grade outcomes and evaluation scores (Quansah et al., 2024).

The extremely high internal consistency ($\alpha \approx .98$) suggests that the instrument functions as a near-unidimensional measure. Inter-item correlations indicate that students respond globally rather than discriminating sharply among constructs such as clarity, engagement, and organization.

This supports the argument that SET instruments often capture overall affective impressions rather than discrete pedagogical dimensions. When internal consistency approaches redundancy levels, interpreting individual items as separate constructs becomes questionable. Institutions that treat each item independently for performance evaluation may be overestimating the instrument's dimensional precision.

Although grade data were not analyzed in direct relation to SET outcomes in this study, grading structures and workload perceptions may differ across modalities and influence student perceptions. Empirical research further demonstrates that student ratings do not always align with objective learning outcomes. Boring (2017) found that evaluation scores diverged from exam-based performance measures, suggesting that perceived teaching quality and measurable learning may operate independently. Stroebe (2016) similarly cautioned that institutional reliance on SET for personnel decisions may incentivize grading leniency, thereby intertwining evaluation outcomes with grade satisfaction. The present findings should therefore be understood within a framework that recognizes these documented limitations.

SET and Academic Outcomes

Composite SET scores demonstrated small but statistically significant positive correlations with GPA and the percentage of A and B grades. These associations may reflect multiple potential mechanisms. One possibility is that effective instruction contributes to stronger academic performance, which in turn results in higher student evaluations. Alternatively, grade satisfaction may influence students' perceptions of instructor effectiveness, leading to more favorable evaluations independent of objective learning gains.

A reciprocal relationship is also plausible, in which learning success and positive evaluation reinforce one another over time. Because the data are correlational, causal direction

cannot be determined. However, the moderate strength of the observed associations is consistent with prior literature suggesting that grades and evaluations are related constructs but are not interchangeable indicators of instructional quality. These findings underscore the importance of interpreting SET results within a broader instructional and evaluative context rather than treating them as direct proxies for student learning outcomes.

Implications for Institutional Practice

Taken together, the findings of this study align with a substantial body of literature indicating that student evaluations are influenced by structural, contextual, and perceptual factors beyond instructional competence alone (Morley, 2014; Quansah et al., 2024; Stroebe, 2016). The small magnitude of the observed modality difference suggests that institutional interpretation should emphasize contextual understanding rather than categorical conclusions regarding delivery format. As such, direct cross-modality comparisons without contextual consideration may risk attributing structural differences in course delivery to differences in instructional quality. Small numerical distinctions, particularly those below one-tenth of a scale point, may not represent substantively meaningful variation and therefore should be interpreted cautiously in personnel or evaluative decisions.

Institutional decision-makers should also consider the potential influence of response rate variability across modalities. Differences in student participation, particularly in online courses where evaluation completion rates may fluctuate, can affect the representativeness of aggregated scores and warrant careful contextual interpretation. In addition, the near-unidimensional structure observed in the evaluation instrument suggests that composite score interpretation may be more defensible than isolated scrutiny of individual items. When items function cohesively as

indicators of a general evaluation construct, excessive emphasis on single-question differences may introduce interpretive instability.

Consistent with broader SET scholarship, student evaluations should not function as stand-alone indicators of instructional quality in high-stakes decisions but may be more defensible when used alongside complementary measures of teaching effectiveness (Quansah et al., 2024). Multi-measure evaluation systems that incorporate peer review, self-reflection, course design analysis, and evidence of student learning provide a more comprehensive and balanced approach to instructional assessment. Finally, professional development efforts may benefit from attention to elements that students consistently associate with effective instruction, particularly in online contexts where communication clarity, structured organization, and timely feedback play a central role in shaping student perceptions.

Limitations

Several limitations must be acknowledged. First, the dataset was derived from a single institution using one standardized evaluation instrument. Generalizability to other institutional contexts, disciplines, or evaluation designs is limited. Second, response rates were not analyzed in depth across modalities. Differential participation could introduce nonresponse bias, particularly in online environments where response patterns may vary. Chapman and Joines (2017) noted that response rates in online systems can vary substantially and may influence the representativeness of evaluation data. Third, the study relied exclusively on aggregated SET scores and did not incorporate direct measures of teaching practices, student engagement behaviors, or learning outcomes. Consequently, conclusions regarding instructional effectiveness cannot be drawn from rating differences alone.

Recommendations for Future Research

Future research should expand upon the present findings in several important ways. First, incorporating additional explanatory variables such as course level, academic discipline, class size, instructor experience, and student demographic characteristics would allow for a more comprehensive analysis of whether observed modality effects persist after accounting for these contextual factors. Including such variables may help clarify whether differences in student evaluations are attributable to delivery format itself or to structural and instructional characteristics associated with specific course types.

Second, future investigations should examine response patterns and the potential influence of nonresponse bias across modalities. Because student participation in end-of-course evaluations is voluntary, variations in response rates may affect the representativeness of aggregated scores. A more detailed analysis of participation trends could strengthen understanding of how modality interacts with evaluation completion behavior.

Third, integrating objective measures of student learning alongside SET outcomes would provide a more robust framework for interpreting modality-related rating differences. Examining whether differences in evaluation scores correspond to measurable differences in student achievement could help clarify the relationship between perceived instructional quality and demonstrated learning outcomes.

In addition, qualitative analysis of open-ended student comments may offer valuable insight into how students conceptualize effective teaching in online versus face-to-face contexts. Exploring narrative feedback could illuminate themes related to communication, instructor presence, course organization, and engagement that are not fully captured by scaled survey items.

Finally, future research may benefit from investigating alternative or revised evaluation instruments that emphasize observable teaching behaviors and reduce susceptibility to halo effects and other perceptual biases. Continued examination of instrument design may contribute to more valid and context-sensitive approaches to evaluating teaching effectiveness across instructional modalities.

Conclusion

This study identified a statistically significant but small difference in SET scores between face-to-face and online sections. When interpreted within the broader SET literature, these findings reinforce longstanding concerns regarding the interpretation of small mean differences, the distinction between reliability and validity, and the influence of contextual factors on student ratings. Although the instrument demonstrated strong internal consistency reliability, prior research cautions that reliability alone does not establish validity or support high-stakes decision-making (Morely, 2014; Quansah et al., 2024).

The results do not suggest that one instructional modality is inherently superior to another. Rather, they indicate that modality may influence student perceptions in subtle ways that require careful contextual interpretation. Institutions should therefore employ SET data as one component within a broader, multi-faceted framework for evaluating teaching effectiveness.

Appendix A

The 12 SET items are as follows:

1. My instructor helped me engage with the course material.
2. My instructor was well prepared for each class.
3. My instructor communicated ideas and concepts clearly.
4. My instructor organized this course well.
5. My instructor demonstrated knowledge of the subject.
6. My instructor facilitated productive class discussions and activities.
7. The feedback from my instructor assisted my learning.
8. My instructor was regularly available for consultation.
9. My instructor encouraged an atmosphere of mutual respect.
10. My instructor helped me feel that I belonged in this course.
11. Assignments helped me to learn the course content and skills.
12. Course materials were useful in my learning and relevant to this course.

References

Aldeman, M. R., Calkins, C. M., & Branoff, T. J. (2025). The impact of instructional modality on student course evaluations: a comparative analysis. *Educational Research and Evaluation An International Journal on Theory and Practice, 30*, 320-343. https://doi.org/10.1080/13803611.2024.2434594

Anderson, R., & Soden, R. (2023). Improving response rates in online course evaluations: An experimental approach. *Assessment & Evaluation in Higher Education, 48*(3), 378-392.

Aragon, S. R., & Davis, M. (2022). The adjunct effect: Evaluating instructional legitimacy in online education. *Online Learning Journal, 26*(1), 45-63.

Arias, J. J., Swinton, J., & Anderson, K. (2018). Online vs. face-to-face: A comparison of student outcomes with random assignment. *e-Journal of Business Education & Scholarship of Teaching, 12*(2), 1-23.

Avery, C., Bryant, P., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *Journal of Economic Education, 37*(1), 21-37.

Basow, S. A., & Martin, J. L. (2012). Bias in student evaluations. In *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 40-49). Society for the Teaching of Psychology.

Bavishi, A., Madera, J. M., & Hebl, M. R. (2010). The effect of professor ethnicity and gender on student evaluations: Judged before met. *Journal of Diversity in Higher Education, 3*(4), 245-256. https://doi.org/10.1037/a0020763

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

Benton, S. L., & Li, D. (2015a). IDEA research report: Online versus traditional course evaluations. *IDEA Research Report #8*. https://www.ideaedu.org/Research/IDEA-Papers-Series

Benton, S. L., & Li, D. (2015b). Student evaluation of teaching in web-based classes. *IDEA Research Report, 21*, 1-16.

Beran, T., & Rokosh, J. (2009). The consequential validity of student ratings: What do instructors really think? *Canadian Journal of Higher Education, 39*(2), 61-81.

Beránek, M., & Zvára, K. (2022). Longitudinal reliability and context effects in student evaluations of teaching. *Assessment & Evaluation in Higher Education, 47*(4), 635-650. https://doi.org/10.1080/02602938.2021.1902410

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.

Berk, R. A. (2022). Enhancing student participation in online course evaluations: Evidence-based strategies. *College Teaching, 70*(1), 45-53.

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

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, 1-11. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Borup, J., Graham, C. R., West, R. E., Archambault, L., & Spring, K. (2021). The adolescent community of engagement framework: Supporting online student engagement. *Journal of Technology and Teacher Education, 29*, 5-28.

Boysen, G. A. (2015). Significant interpretation and the evaluation of teaching: Implications of a meta-analysis. *College Teaching, 62*(2), 74-79.

Bunn, M., Fischer, J., & Marsh, M. (2021). Comparing student evaluations across modalities: Evidence from a large-scale paired-section study. *Journal of Educational Measurement and Evaluation, 63*(2), 115-134.

Burgoon, J. K. (1993). Interpersonal expectations, expectancy violations, and emotional communication. *Journal of Language and Social Psychology, 12*, 30-48. https://doi.org/10.1177/0261927X93121003

Carnegie Classification of Institutions of Higher Education. (2022). *Basic classification description*. https://carnegieclassifications.acenet.edu

Chapman, D. D., & Joines, J. A. (2017). Strategies for increasing response rates for online end-of-course evaluations. *International Journal of Teaching and Learning in Higher Education, 29*(1), 47-60.

Chavez, K., & Mitchell, K. (2021). Unmasking racial bias in SET: A critical analysis of faculty of color experiences. *Journal of Diversity in Higher Education, 14*(3), 299-311.

Chen, L., & Schunn, C. D. (2025). When student evaluation of teaching is no longer indicative: Psychology students' in-class experience, motivational factors, and future choices. *Studies in Educational Evaluation, 86*, 1-11. https://doi.org/10.1016/j.stueduc.2025.101475

Chen, Y., & Hoshower, L. B. (2021). Student motivation to participate in online evaluations: A revised model of SET response behavior. *Research in Higher Education, 62*(5), 672-691.

Clayson, D. E. (2021). Toward a deeper understanding of engagement in the classroom: How and why it matters. *Journal of Marketing Education, 43*(1), 28-44. https://doi.org/10.1177/0273475320929286

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*(3), 281-309. https://doi.org/10.2307/1170209

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research, 41*(5), 511-535. https://doi.org/10.2307/1169890

Crimmins, G., Casey, S., Weber, I., & Pourfakhimi, S. (2024). Consideration of student evaluations of teaching (SET) and learning: Perspectives of learning and teaching leaders through the ethical lens of 'first, do no harm'. *Assessment & Evaluation in Higher Education, 50*(1). https://doi.org/10.1080/02602938.2024.2367586

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

Daskalopoulou, M. (2024). Gendered experiences of evaluation in academia: A critical review. *Higher Education Review, 56*(1), 45-62.

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

Eagly, A. H. (1987). Sex differences in social behavior: A social-role interpretation. *Erlbaum*.

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

Flaherty, C. (2023). *Student evaluations and faculty well-being: New data on the harm of abusive comments*. https://www.insidehighered.com/news/2023/01/25/new-research-examines-student-evaluation-abuse

Fynn, A. (2024). *Validity of student evaluation of teaching in higher education: A systematic review*. https://www.example.com

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

Gupta, A., & Henricksen, R. (2024). Beyond the platform: Course design as a mediator of teaching evaluations in online and in-person courses. *International Review of Research in Open and Distributed Learning, 25*(1), 1-17.

Heffernan, A. (2022). Gendered experiences in student evaluations of teaching: A systematic review. *Teaching in Higher Education, 27*(2), 157-173.

Heffernan, T. (2022). The tyranny of student evaluation of teaching. *Journal of Higher Education Policy and Management, 44*(2), 1-15. https://doi.org/https://doi.org/10.1080/1360080X.2021.1957151

Heffernan, T. (2023). 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

Heffernan, T. A. (2022). The performance penalty for women in academia: A review of gender bias in student evaluations of teaching. *Assessment & Evaluation in Higher Education, 47*(3), 403-416. https://doi.org/10.1080/02602938.2021.1946745

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

Jabeen, S., Saeed, A., & Kumari, N. (2024). Examining the relationship between student evaluation of teaching, instructional quality, and faculty development in Pakistani higher education. *Bulletin of Education and Research, 46*(2), 15-34.

Johnson, M. (2020). *Online and face-to-face teaching: How do student ratings differ?* https://www.example.com

Johnson, R. (2020). The impact of delivery mode on student evaluations of teaching. *Journal of Online Learning Research, 6*(1), 1-16.

Kember, D., Leung, D. Y., & Kwan, K. P. (2023). The use of generalizability theory in interpreting student feedback: Addressing instability in low-response scenarios. *Quality Assurance in Education, 31*(2), 210-225.

Kreitzer, R., & Sweet-Cushman, J. (2021). Evaluating student evaluations of teaching: A review of measurement and equity bias in SETs and recommendations for ethical reform. *Journal of Academic Ethics*. https://doi.org/10.1007/s10805-021-09400-w

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

Lewis, K., & Freeman, R. (2023). Students' perception of clarity and its role in overall satisfaction: A comparative study of online and traditional classrooms. *Teaching in Higher Education, 28*(2), 215-230.

Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation, 54*, 94-106. https://doi.org/10.1016/j.stueduc.2016.12.004

Liu, R., & Carless, D. (2023). Reconsidering the representativeness of online student evaluations: A mixed-methods study. *Higher Education Research & Development, 42*(3), 541-559.

Macfayden, L. P., Dawson, S., Pardo, A., & Gašević, D. (2023). Reimagining modality: A multi-institutional analysis of SET across hybrid, online, and face-to-face contexts. *Higher Education Research & Development, 42*(4), 856-873.

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

Marsh, H. W. (1980). The influence of student, course, and instructor characteristics in evaluations of university teaching. *American Educational Research Journal, 17*(2), 219-237. https://doi.org/10.2307/1162484

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. *Journal of Educational Psychology, 76*(5), 707-754. https://doi.org/10.1037/0022-0663.76.5.707

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*(3), 253-388.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187-1197.

Martin, F., & Borup, J. (2022). Interpreting student engagement and feedback in virtual classrooms: A mixed-methods study. *Online Learning, 26*(2), 75-96.

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

Mitchell, K., & Martin, K. (2023). Hostility in the classroom: A study of anonymous student feedback. *Teaching Sociology, 51*(1), 46-61.

Monnier, N. (May-June 2017). 'One faculty' and academic governance: Toward a meaningful role in governance for non-tenure-track faculty. *Academe Magazine*. https://www.aaup.org/academe/issues/103-1/one-faculty-and-academic-governance?utm_source=chatgpt.com

Morely, D. (2014). Assessing the reliability of student evaluations of teaching: choosing the right coefficient. *Assessment & Evaluation in Higher Education, 39*(2), 127-139. https://doi.org/10.1080/02602938.2013.796508

Owens, J., & Zhang, J. (2022). Participation in digital course evaluations: A comparative study of online and in-person modalities. *Online Learning, 26*(4), 85-102.

Papadogiannis, I., Vassilakis, C., Wallace, M., & Katsis, A. (2024). Challenges and trends in student evaluation of teaching: Analysis of SET data from the University of Peloponnese. *Information, 15*(576), 1-16. https://doi.org/10.3390/info15090576

Peterson, R., & Miles, J. (2024). From evaluation to feedback: Reframing student input to promote equity. *Higher Education Policy, 37*(1), 89-105.

Pineda, P., & Steinhardt, I. (2023). The debate on student evaluations of teaching: Global convergence confronts higher education traditions. *Teaching in Higher Education, 28*(4), 859-879. https://doi.org/10.1080/13562517.2020.1863351

Pittman, K. M. (2022). Discriminatory consequences of student evaluations in promotion decisions: A legal analysis. *Journal of Academic Ethics, 20*(2), 131-146.

Qin, Y. (2025). Online vs. face-to-face: A long-term study on the effectiveness and essence of learning. *Cogent Education, 12*(1). https://doi.org/10.1080/2331186X.2025.2554314

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

Richardson, C., Fielding, A., & Ng, J. (2024). Who responds to student evaluations? Demographic predictors and implications for interpretation. *Studies in Educational Evaluation, 81*.

Rivera, J., & Beranek, T. (2023). Online versus face-to-face student evaluations: A meta-analytic review. *Journal of Online Learning Research, 9*(2), 101-126.

Rovai, A. P., Ponton, M. K., Derrick, M. G., & Davis, J. M. (2006). Student evaluation of teaching in face-to-face and online classes. *The Internet and Higher Education, 9*(1), 23-35. https://doi.org/10.1016/j.iheduc.2005.11.002

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.

Senden, B., Nilsen, T., & Teig, N. (2023). The validity of student ratings of teaching quality: Factorial structure, comparability, and the relation to achievement. *Studies in Educational Evaluation, 78*, 1-12. https://doi.org/10.1016/j.stueduc.2023.101274

Sidwell, D., Lee, D., Zimmerman, P., Bentley, S., & Barton, M. (2025). Teaching faculty experiences with student evaluation of instruction: A mixed-methods study. *Teaching and Learning in Nursing, 20*(1). https://doi.org/10.1016/j.teln.2024.11.009

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

Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science, 11*(6), 800-816. https://doi.org/10.1177/1745691616650284

Stroebe, W. (2020a). Student evaluations of teaching encourage poor teaching and contribute to grade inflation: A theoretical and empirical analysis. *Basic and Applied Social Psychology, 42*(4), 276-294. https://doi.org/10.1080/01973533.2020.1756817

Stroebe, W. (2020b). Student evaluations of teaching: Potential biases and their implications for teaching effectiveness. *Educational Psychology Review, 32*, 843-877. https://doi.org/10.1007/s10648-020-09512-5

Sullivan, D., Lakeman, R., Massey, D., Nasrawi, D., Tower, M., & Lee, M. (2024). Student motivations, perceptions and opinions of participating in student evaluation of teaching surveys: A scoping review. *Assessment & Evaluation in Higher Education, 49*(2), 178-189. https://doi.org/10.1080/02602938.2023.2199486

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tomas, N., Aukelo, M., & Tomas, T. N. (2022). Factors influencing undergraduate nursing students' evaluation of teaching effectiveness in a nursing rogram at a higher education institution in Namibia. *International Journal of Africa Nursing Sciences, 17*, 1-6. https://doi.org/10.1016/j.ijans.2022.100494

Tu, C. Z., & Adkins, J. (2020). Comparing students' evaluation of online, hybrid, and face-to-face classes. *Assessment & Evaluation in Higher Education, 45*(6), 97-100. https://doi.org/10.54808/ICSIT2022.01.97

U.S. Department of Education. (2023). The condition of education 2023: Postsecondary institutions. *National Center for Education Statistics*. https://nces.ed.gov

UNESCO. (2023). International Standard Classification of Education (ISCED) 2011: Levels of education and fields of training. *United Nations Educational, Scientific and Cultural Organization*.

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

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22-42.

Wang, N., & Zhao, Z. (2025). SuperHyperSoft Set-based evaluation of teaching quality in ideological and political courses in colleges and universities in the new era. *Neutrosophic Sets and Systems, 85*, 472-487.

Yamada, R., & Sato, Y. (2021). Examining the influence of SET on faculty promotion decisions: A longitudinal analysis. *Journal of Higher Education Policy and Management, 43*(2), 155-170.

Young, S., & Duncan, H. (2014). Online and face-to-face teaching: How do student ratings differ. *MERLOT Journal of Online Learning and Teaching, 10*(1), 70-79.

Zhao, J., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: Exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education, 37*(2), 227–235. https://doi.org/10.1080/02602938.2010.523819

Zhou, Y., & Sloan, T. (2022). The accent effect: Student perceptions of international faculty in higher education. *International Journal of Teaching and Learning in Higher Education, 34*(1), 99-109.