

University of Texas Rio Grande Valley

**ScholarWorks @ UTRGV**

---

Criminal Justice Faculty Publications and  
Presentations

College of Liberal Arts

---

6-2023

## **Student Evaluations of Teaching Are Mostly Awfully Wrong**

Noel Otu

Ntiense E. Otu

Follow this and additional works at: [https://scholarworks.utrgv.edu/cj\\_fac](https://scholarworks.utrgv.edu/cj_fac)



Part of the [Criminology and Criminal Justice Commons](#), and the [Higher Education Commons](#)

---

---

## STUDENT EVALUATIONS OF TEACHING ARE MOSTLY AWFULLY WRONG

Noel Otu<sup>1\*</sup>, Ntiense E. Otu<sup>2</sup>

The University of Texas Rio Grande Valley, Brownsville, Texas, USA<sup>1</sup>  
South Texas College, McAllen, Texas, USA<sup>2</sup>

\*Email: [noel.otu@utrgv.edu](mailto:noel.otu@utrgv.edu)

---

Received: April 2023

Accepted: May 2023

Published: June 2023

### ABSTRACT

*Student evaluations of teaching (SETs) have been used, researched, and debated for many decades. It is a common practice in higher education institutions, with the supposed purpose of improving course quality and effectiveness, but with unintended consequences of encouraging and motivating poor teaching and causing grade inflation. There is strong evidence that SET “effectiveness” does not measure teaching effectiveness. This paper reviews empirical research examining common concerns about the usefulness (positive and negative) and accuracy of SETs. The findings reveal that student satisfaction relates to their anticipated/expected grades in their courses; hence, they want to get good grades and their instructors want to get a good rating of SET, and this results in grade inflation. The key points are that SETs (1) allow students to speak their “mind”, (2) have no compelling correlation between quality of teaching and learning effectiveness, (3) reward easy, less demanding, and lazy teachers with a positive rating, (4) are biased against gender, attractiveness, ethnicity, race, etc., (5) are weaponized against “some” faculty members, and (6) are like asking convicts awaiting sentencing to evaluate the judge or jurors who convicted them.*

**Keywords:** Student evaluation of teaching, tenure and promotion, retaliation, biases

### Suggested citation:

Otu, N., & Otu, N. E. (2023). Student Evaluations of Teaching Are Mostly Awfully Wrong. *Universal Journal of Educational Research*, 2(2), 168-183.

## INTRODUCTION

A student evaluation of teaching (SET) is a questionnaire or survey, typically administered by hand copy or with an online evaluation form, designed to evaluate a given course at a higher education institution (Himmelsback, 2019). The SET originated in the early 20th century, and its use spread worldwide in the 1960s and 1970s (Remmers and Brandenburg, 1927; Marsh, 1987; Remmers, 1933). This was because students wanted a voice at their universities. The aim of SETs is to provide feedback and insight from students about a particular course that will help the instructors identify what's working in the delivery and content of a course and to determine what could be improved, from content, to required activities, to delivery style (Himmelsback, 2019).

The evaluation results were primarily used to help faculty improve their teaching until the 1970s, when administrators began applying them to personnel decisions (Ward, 2021, p 2; Galbraith, Merrill, and Kline, 2012). According to Clayson (2021, p 3), the purpose of the evaluation in the 1960s and 1970s varied widely, with "no consensus about what the process was designed to measure." Also, the use of SETs developed from an increasing emphasis on "quality assurance," "performance management," and "consumer satisfaction" in institutions of higher learning (Spooren, Brockx, and Mortelmans, 2013). Additionally, Cox, Rickard, and Lowery (2021) state that the high importance of SETs in higher education is the result of institutions' adoption of consumerism principles. These principles lead students to pressure faculty for higher grades and to threaten to give bad ratings if the faculty fail to give these grades (Ward, 2021). It has been acknowledged that student evaluations have grown in use and power as a measure of "accountability agenda," which has resulted in the "devaluing of faculty" (Titus, 2008, p 413–414; Ward, 2021).

There are two sides to the debate for and against SETs. One side of the argument assumes that scores students give instructors are moderately correlated with student learning and the best pedagogical practices (Esarey et al, 2020; Huemer, 2001, p 1). Some studies assume that the SETs are highly reliable, or that instructors consistently get the same ratings. It also supposes that SETs do not systematically discriminate on any instructionally relevant basis (Esarey and Valdes, 2020).

SETs are highly reliable in that students tend to agree with each other in their ratings of an instructor (Huemer, 2001, p 1), and they are at least moderately valid. They are often useful because instructors who receive mid-semester evaluations tend to have a higher rating in end-of-semester evaluations than those who do not, suggesting that SET results cause changes in teaching behavior, which results in higher ratings (Huemer 2001, p 1). Although the improvement could be attributed to just about any factor, including, but not limited to, grading leniency and/or dumbing/watering down courses. A particular simulation study is the only one that says anything even "moderately" fair about SETs, and the authors immediately turned around in the same article and declared them to be flawed (Esarey and Valdes, 2020). "Steadily accumulating evidence of the misuse or overuse of rating data – and the perennial debate in the press concerning the validity of student ratings – do not invalidate the potential of rating data as useful information about teaching performance" (Theall and Fracklin, 2000, p 95).

On the other side of the debate about the validity and fairness of SETs, several studies state that SETs actually undermine learning by encouraging lenient grading and superficial classroom presentation (Johnson 2003; Esarey and Valdes, 2020). Extensive research by psychologists and educators convincingly demonstrated that SETs are biased, unfair, invalid, and flawed (Merritt, 2008). Psychology research demonstrated that the human mind functions along two very different tracks, one that generates automatic, instinctive sanctions and another that produces more reflective, deliberative decisions (Sloman, 2002). The way that we currently obtain teaching evaluations from students taps into instinctive rather than selective judgments (Merritt, 2008). When stacking the deck of SETs, research finds that they are deeply flawed (Esarey and Valdes, 2020).

The most common criticism of SETs seems to be that they are biased in that students tend to give higher ratings when they expect higher grades in their course. The correlation is well established (Huemer, 2001). SETs are most likely a reflection of a student's satisfaction with a course, which can be influenced by factors unrelated to teaching effectiveness (Fristitit, 2016). Hence, the use of SETs as a basis for

decisions on tenure and promotion would be justified if SETs were a valid measure of teaching effectiveness and student learning, which they are not (Boring, 2017; Uttl and Smibert, 2017; Stroebe, 2020).

The aim of this paper is to provide evidence of the impact of other factors outside effective teaching and learning that influence and determine SET ratings. Also, basic assumptions, including many types of biases, but not limited to grading leniency, will be presented. The bias assumption states that the work students are required to invest in a course and the grades they expect to get influence their evaluation of courses and instructors (Stroebe, 2016). Other biases are race, gender, and appearance. Minority faculty receive significantly lower evaluations than their white colleagues (Harmensiesh and Parker, 2005). According to Merritt (2008), white faculty members have also noted the possibility of bias in their student evaluations, particularly based on gender, appearance, and political ideology (Abel, 1990; Beam, 1989). Just as with individual differences in human beings, each university and/or department is structured by caste and has unique features. In the academic field, judgments of worth tend to be very subjective. Hence, academic decisions are very susceptible to unconscious or conscious biases, and they take different forms from one university/department to another, by individuals or collectively.

It is fair to both students and faculty to critically study the SET system that we currently use, which is not working well. Revealing the biases in our SET process and the unintended consequences that follow will help eliminate the flaws in the system. Appreciating the biases involved will best explain a good way of measuring teaching effectiveness. This article begins the pathway to creating a common knowledge of what SETs do and do not measure.

### **Statement of the Problem**

The common misconception is that SETs in and of higher education are a balancing act between control, public accountability, and quality improvement (Dabler-Lerren, 2009; Dano and Stensaker, 2007; Raban, 2007; Williams, 2016; Borch, Sandvoll, and Riser, 2020). In reality and practice, SETs can be utilized as institutionally “legitimized” weapons to bully faculty up for personnel action – tenure, promotion, etc. (Rodriguez, 2019). Seldin (1993) states that there is an 86 percent use of SETs as a central feature of personnel decisions in US higher education, an increase in usage from 68 percent in 1984 and 28 percent in 1973 (Seldin, 1984; Pounder, 2007).

Across many university campuses, there are growing ideas that frontline workers in higher education should “serve their customers like those in other industries.” Many universities have adopted a retail industry approach. The major reason why students have developed a retail-service mindset is because colleges and universities have successfully promoted that mentality as a means to attract and retain students (Searey, 2017). For the past 25 years, the retail business model has become the de facto one for higher education. “Specialization of the retail business model is reflected in the institutionalization of enrollment management. Admissions, student affair, registration, financial aid and the bursar – once siloed – now coalesce in enrollment management” (Searay, 2017, p 1). It is common knowledge to professors that legislators, accrediting bodies, alumni, donors, and even university administration see retention and graduation rates, along with job attainment, as assessments of institutional effectiveness. Enrollment management has become one of the most important parts of university administration as a result and customer service is viewed as a priority (Searey, 2017).

With the customer service mentality and discrimination, SETs can be weaponized to justify undermining academic freedom and subjecting racial minorities, women, and untenured professors to surveillance and bullying (Rodriguez, 2019; Merritt, 2008). Professors of color have published poignant accounts of harshly negative student evaluations (Smith, 1999). In today’s world, race and gender remain supremely relevant to whether some professors are perceived as too incompetent to profess. The negative characterizations of women of color are exacerbated when black women exercise authority in academia, for example, through class management, grading, and other forms of student evaluative measures (Smith, 1999). There are empirical studies examining instructor race, and student ratings have confirmed that minority faculty receive significantly lower evaluations than their white colleagues (Hamermesh and Parker, 2005). White faculty members have also noted the possibility of bias in their student evaluations, particularly based on gender and appearance (Merritt, 2008). Several studies have demonstrated that

pregnancy is viewed as a lifestyle choice (Sieverding, Elb, Neubauer, and Stahl, 2018; Baker and Copp, 1997), while illness is beyond one's control. Women who "choose" to get pregnant, particularly women of color, may be considered "decreasing" of poorer teaching evaluations (Olabisi, 2021).

While bullying at workplaces generally involves all races and genders, it is particularly pertinacious for the underrepresented populations in academia. Despite anti-discrimination legislation, affirmative action initiatives, and higher numbers of black students graduating with doctorates, African American faculty are underrepresented in colleges and universities (Allen, Epps, Guillory, Suh, and Bonous-Hammarth, 2000; Frazier, 2011). The current representation of faculty of color is very low and one may believe the reason is due to universities focusing on the recruitment rather than the retention of African American faculty (Frazier, 2011; Trower and Chait, 2002). Studies have shown that people of color and women experience bullying to a far greater degree than their white male counterparts, and this bullying occurs more often in higher education than in other workplaces (Swann, 2022; Hollis, 2022). African American faculty, unfortunately, have been forced into a revolving door scenario in academia. Often, they are brought into majority white institutions and become victims of a systematically racist institutional climate and feelings of isolation (Frazier 2011; Patity and Hinton, 2003; Tillman, 2001).

### Review of Literature

This study enters a literature that is both large and divided in its assessment of the value of SETs, but perfectly united in that its focus is on the use, misuse, and abuse of SET scores as a tool for harassing, intimidating, and bullying minorities and women faculty members. In favor of SETs, some studies report that student ratings of instruction produced by IDEA "correlate positively with external student ratings of learner and teacher behaviors" as well as student performance in tests (Benton, Duchon, and Pallett, 2013; Esarey and Valdes, 2020). SET ratings somehow break into multiple dimensions that correspond to student proficiencies (Esarey and Valdes, 2020; Li, Benton, Brown, Sullivan, and Ryalls, 2016). According to Esarey et al (2020), "SETs are moderately correlated with teaching quality (student learning and instructional best practices), highly reliable, and do not systematically discriminate on any instructionally irrelevant basis" (p 1). Huemer (2001) states that "SETs are highly reliable, in that students tend to agree with each other in their ratings of an instructor and that they are at least moderately valid, in that student ratings of course quality correlate positively with other measures of teaching effectiveness" (p 1). SETs tend to correlate well with retrospective evaluations by alumni (Huemer, 2001). It is common knowledge to teachers that students take many courses and should therefore be able to judge whether a particular teacher is effective. This assumption is supported by the fact that course evaluations are positively correlated with the grades the students received in those courses (Stroebe, 2020).

On the other side of the SET debate, many researchers have argued that SETs do not measure student learning or instructor quality and may be systematically biased against some groups, irrespective of merit (Esarey et al, 2020; Allen, Epps, Guillory, Sun, and Bonous-Hommarth, 2000). In a meta-analysis of research from the 1950s to the mid-2000s, studies found little or no link between learning and student surveys of teaching and suggested that rigor in a class led to lower ratings in surveys (Clayson, 2021). It is common knowledge to teachers that students who perform well in class will give higher scores to instructors than students who do not (Clayson, 2021).

Clayson (2021) states that administrators should not use scores from student surveys as a method of comparing instructor performance. Also, Onwuegbuzie, Daniel, and Collins (2009) concluded that their study of multiple components of validity in students' surveys for evaluating teaching raises "serious doubt" about the use of them for this purpose. According to Spooren, Brocky, and Mortelmans' (2013) analysis of 12 years of scholarship, studies have failed to provide clear answers about the validity of students' surveys to evaluate teaching, and the tension between their use as formative and evaluative tools makes them fragile (p 599). They went on to say that a new approach may be needed because of a shift toward a student-centered (consumerism) teaching mentality. In a more recent study, Esarey and Valdes (2020) declared that:

"Using invalid, unreliable or biased student evaluations to make decisions about hiring and tenure is obviously harmful to students and faculty alike. When universities use SETs that

do not track student learning or instructional best practices as a part of hiring, tenure and promotion process, they may make important decision on the basis of random chance instead of merit.” (p 7)

The existing literature around SETs is rarely positive, and while this paper primarily deals with the bias and invalidity issues of SET ratings, it is important to note that this research contributes to the prevailing work, bringing together different and multiple problematic natures of use of student evaluation. It is also important to show both sides of the argument – positive and negative – of SETs ratings. Multiple methodological flaws reduce the validity of SETs as a measure of teachers’ effectiveness (Rampel, 2018). First, students evaluate partial courses: schools typically solicit student evaluations before completion of a course. Second, sample sizes have shrunk: schools’ replacement of in-class, pencil-and-paper exercises with cheaper, take-home, online response systems has been accompanied by a drop of 30 percent or more in response rate (Rampel, 2018; Jan, 2010). Hence, SETs are problematic.

## METHODOLOGY

This paper has selected qualitative data collection and analysis methods as these allow a systematic approach to the human element and lived experiences of, in this case, academics on the receiving end of abuse of SETs (Hefferman, 2022). According to Shrilman (1988), it is the focus of the research questions needing to be answered that dictates the methods that should be used. With this in mind, Denzin and Lincoln (2011) concluded that it is the responsibility of the researchers to ensure they have the full array of research methods available to them (Hefferman, 2022).

The qualitative data collection technique used is called triangulation, which is a combination of different methods used to explore research hypotheses. The methodology was tailored to gain as much detailed and multilayered information as possible about SET validity, invalidity, and different types of biases associated with this evaluation. It followed Kopinak’s definition of method triangulation as entailing “gathering information pertaining to the same phenomenon through more than one method, primarily in order to determine if there is a convergence and hence increased validity in research findings” (Kopinak, 1999, p 171). Several issues were of importance in the use of triangulation in this study, including the sampling of research materials and thorough analysis of information. It is important to note that the two researchers who conducted this study have a combined 35 years of college-level teaching history at five different colleges and have experienced about 70 SETs. In addition to numerous literatures reviewed for this study and the lengthy number of years of researchers’ experiences in the subject matter, the use of triangulation enhances validity because it involves the understanding that many factors cause bias in SETs, and it is therefore beneficial to this research because it allows us to look at the pros and cons in literatures about SET and to manipulate/analyze more than one factor (Mason, 2010).

### Analysis of Data

The data consist of reviews of numerous literatures and transcripts from manuscripts, books, and microfilms related to SETs. According to Kellehear (1993), the three general forms of analysis in qualitative research are a) semiotic discourse and reflective analysis, b) thematic analysis, and c) content analysis.

The data analysis of this study is done inductively and deductively (Fook, Ryan, and Hawkin, 2000). The inductive aspect of the study is comprised of a thematic analysis, which involves analyzing data for the bias and invalidity in SETs, with emphasis on how they undermine learning by encouraging lenient grading and superficial classroom presentation (Johnson, 2003). The deductive part of the study is comprised of a content analysis of how race, gender, appearance, etc. influence SETs. The deductive analysis is interested in linking/comparing this study with others (for example, Smith, 1999; Merritt, 2003; Clayson, 2021; Esarey and Valdes, 2020), which academically uplifts the integrity of this research. The major difference between this research and others is that it looked at the good, the bad, and the ugly sides

of SETs. The deductive analysis was used as part of the broader inductive analysis in order to develop a more complete, complex, and clear picture that SETs are biased. The repetitive process of analyzing the triangulated data generated by the analysis of books and articles yielded major findings. These findings center around the fact that student teaching evaluation allows irrelevant characteristics like race, gender, appearance, etc. to distort SETs. To straightforwardly interpret the findings, we will first provide a short definition of SETs.

### **SETs Defined**

This research uses the term SET to refer to the end-of-semester surveys that gather student feedback about university courses and instructors. These surveys are sometimes referred to as student ratings, course evaluation, and student surveys of teaching (Ward, 2021). The original aim of SETs was to help faculty improve their teaching. It was during the 1970s that administrators began using them for personnel decisions (Galbraith, Merrill, and Kline, 2012; Ward, 2021). According to Clayson (2021, p 3), “the purpose of SET[s] in the 1960s and 1970s varied widely, with no consensus about what the process was designed to measure.” As described by Spooren et al (2013), the use of SETs for faculty evaluation arose from an increasing emphasis on “quality assurance,” “performance management,” and “consumer satisfaction” in universities (Ward, 2021). In addition, Cox, Richard, and Lowery (2021) declare that the use of SETs as an important tool in higher education is the result of universities’ adoption of a consumerist model.

### **SETs vs Consumerism Mentality**

The notion of human beings as consumers first took shape before World War I but became commonplace in America in the 1920s. Over the course of the 20th century, capitalism preserved its momentum by molding the ordinary person into a consumer with an unquenchable thirst for more stuff (Higgs, 2022). Given that the cost of higher education tuition in the United States has continued to rise over the past two decades, and with the impact of COVID-19 changing delivery methods of education, many university administrators expect the best from faculty members, and many students want to dictate what they want (A grades) for their money. According to Edmunson (1997), universities cater to students to boost enrollment and retention numbers, and increase marketing efforts to recruit students (Burke, Hughey, and Cannonier, 2019), as college student customers have produced a “customer is always right” mentality. In the process of finding out the “customer’s” feelings about courses and faculty, students are allowed to complete SET forms that cause them to focus on what they want (easy classes, highest grades, etc.) as opposed to what they potentially need (educational knowledge) (Burke et al, 2019). In higher education, it is very easy to identify customers, consumers, and products, depending on the knowledge level of the individual.

According to Searcy (2017), universities evolved from constitutional provisions that addressed overall education; for example, California’s Article IX, Section I states that its higher education legislation encourages the “promotion of intellectual, scientific, moral and agricultural improvement through ... a general diffusion of knowledge and intelligence being essential to the preservation of the rights and liberties of the people” (Searcy, 2017, p 3). The Constitution of Connecticut, Article 8, Section 2, states: “The state shall maintain a system of higher education, including the University of Connecticut which shall be dedicated to excellence in higher education.” Section 3 establishes the charter for Yale University. Article X, Section 1 of the Illinois constitution allows the state to provide for an efficient system of high-quality public educational institutions and services. Also, the Texas Constitution, Article 7, Section 10, states: “the legislature shall as soon as practicable establish, organize and provide for the maintenance, support and direction of a university of the first class, to be located by a vote of the people of the state, and styled, ‘The University of Texas’ for the promotion of Literature, and Arts and Sciences, including an agricultural, and mechanical departments.”

It is common knowledge to interested parties that since state universities and some private ones were established by state constitutions and are supported by tax dollars, the actual customers of higher education are the citizens of respective states (Searcy, 2017, p 3). The question that we need to answer is: would students’ school fees/tuition build and maintain universities? The answer is categorically *no*.

Customers pay for a product or service; in this case, the government/taxpayers do, but students consuming the services compensate the government with a small amount of money in addition to what the government/taxpayers pay for the service/product the university supplies. Students are consumers, not customers (Searcy, 2017). Consumers use products or services, and since higher education was created and maintained to produce an informed citizenry for a democratic society, being the customers, students are clearly the consumers of the product (knowledge).

Universities generate knowledge and craft tools to assess the mastery of information. There is confusion among some students, faculty, and administrators about applying the students as customers paradigm vs society as customers paradigm to higher education (Searcy, 2017). If students are the customers, then the retail model of “customer is always right” sets in. Customer service determines that all customers should be happy because a good/happy customer is a repeat customer, telling their friends and giving big tips. It is common knowledge to faculty that an “A” grade makes a happy student; happy students give high ratings in SETs.

Although not all university students exhibit these characteristics and claim academic entitlement, those who do can directly create negative consequences for faculty in SETs. According to Chowning and Campbell (2009), SETs are best predicted by students’ academic performance, grades, and their level of academic entitlement. When students believe they should be treated as customers rather than as students (consumers), they believe they are entitled to positive academic outcomes, irrespective of performance, achievement, effort, or personal responsibility, and blurring the boundary lines of the traditional faculty–student relationship (Burke et al, 2019, p 38; Lippmann, Bulanda, and Wagenaar, 2009).

Landrum (1999, p 126) reports that “students performing average work, who acknowledge themselves that their work is average, expect a grade of B or A more than 70 percent of the time, even though they realize that the grade for average work is C.” When more students in a class have these expectations or “gang up” on a faculty member, faculty efforts to curb inflation and evaluate students according to their performance and ability suffer. As the number of students with a customer entitlement mentality increases in the classroom, there is greater pressure on faculty to alter the approved normative grading standard, especially if there is no support from administration that faculty should maintain traditional standards (Lippmann et al, 2009, p 199). According to Pope (2001, p 152), students have devised many strategies and behaviors to pressure faculty so that they achieve their academic success through lower ratings in SETs, filing complaints, contesting grades, ganging up on faculty, and so on. Continual refusal to inflate grades may lead to student–faculty confrontation, resulting from an inability to interact positively with disgruntled students, which may lead to grade inflation just to keep students happy (Johnson, 2003). Finally, inflating grades may lead to the suspension or closure of a program that has an external qualifying examination, like pharmacy, medicine, social work, nursing, etc.

In nursing, it is important that students pass the National Council Licensure Examination (for) Registered Nurses (NCLEX-RN) at the first attempt, since the failure of a program to meet the passing standard – an 80% pass rate – three consecutive times, as set by the Board of Nursing (BON), would result in the closure of the pre-licensure nursing program. The BON has standards expected of an Associate Degree Nursing (AND) program, and it evaluates nursing program effectiveness and outcomes based primarily on first-time NCLEX takers’ pass rates; BON has the authority to suspend or terminate any nursing program that fails to meet the required standards (Czekanski et al, 2018; Denman and Cohn, 2022). According to Norris (2019), “the student most likely to be disadvantaged by the inherent and unavoidable attribute of test item flaw/bias is the learned learner,” since this (nursing) test is a standardized one that requires candidates to choose only one correct response (p 40). It is common knowledge to nursing faculty members that low ratings in SETs may be the consequences of the high standards required of nursing students, and high ratings may be the result of grade inflation. There is a predictive relationship between student rating, expected grade, SET rating, grade inflation, and low rates of NCLEX success. Hence, SETs then become more of an opinion survey, which has never been their intended purpose. When students view faculty merely as a service provider, faculty begin to be impacted and the learning process can become severely compromised (Burke et al, 2019, p 38).



“The customer is always right” is awfully wrong if applicable to students and universities. This phrase, coined in 1909 by British department store owner Harry Gordon Selfridge, was an attempt to convince employees to give great customer service (Face, 2019). Across many universities, the idea of faculty members being considered frontline workers and expected to “serve their customers (students) like those in other industries” has caught up with university administrations, and by doing exactly that, over the past two decades, students have developed a retail service mindset (Searcy, 2017).

Although universities must compete more for fewer traditional students than in the past, some higher education institutions have transitioned to a business model that stresses a bulwark against the tide of decreasing enrollment. The phenomenon of “consumerism” cuts across so many different aspects of university life that it has a direct effect on students’ evaluation of teaching rates. It rewards poor teaching with higher ratings. There is a lack of clarity surrounding customers and products at the higher education level. The product in higher education is knowledge, and mastery of that knowledge is the responsibility of the student who is to consume it. According to Searcy (2017), the fundamental point remains that crafting higher education policy at all levels should be derived from identifying the public, not the student, as the customer in higher education. Clarifying to faculty members that the customer is the public reduces grade inflation pressure, allow faculty to demand excellence for the future of a better society, and rendering SETs worthless.

### SETs vs Reliability and Validity

A “reliable” test is one that gives the same result when repeated and indicates that it must be measuring something. A “valid” test is one that measures what it is intended to measure. Researchers generally consider SETs to be highly reliable in that students tend to agree with each other in their ratings of instructors (Contra, 1993; Hobson and Donna, 2001; Aleamon, 1999). Some researchers also agree that SETs are at least moderately valid (Huemers, 2001; Aleamon, 1999; Hobson and Donna, 2001; Centra, 1993). Evidence is available that suggests there are major issues with the statistical validity of SETs, such as there being non-standard distances between ratings (e.g., an unequal difference in scoring assignment of responses such as excellent, good, fair, poor, or very poor), which makes interpreting results, especially when comparing means, problematic (N/A #1, 2022).

Stark and Frieshtat (2014) state that the variables in SETs are frequently “ordinary category variables,” that is, possible values that have a natural range or order such as “Excellent, Good, Fair, Poor, Very Poor,” but have no standard quantitative distance between each rating. According to N/A # 1 (2022),

“---variables are typically assigned a number, such as 1 to 5, but these numbers are not measurable qualities or quantities, rather, they are only labels from which it is convenient to calculate mean, median, and so forth. One student’s choice of a number on a 1 to 5 scale may not mean the same as another’s, and the context for these choices is not always clear. When the scores are averaged, the precise value (such as a 4.2 in ‘teaching quality’) is difficult to interpret as averages can obscure potentially meaningful variations across responses. For example, the average of two ‘3’ scores is the same as the average of a ‘1’ and a ‘5’, and yet these distributions have very different implications when coming from a group of students. In addition, a ‘bootstrap’ analysis of historical data at Stanford reveals that a variation of +/- 0.3 (more than half a point altogether) can be expected due to random selection from the pool of available students alone. Finally, a significant portion of instructors will always fall below the average, regardless of their individual ability or commitment, if a departmental average is taken as a target for performance.” (p 3)

There is no compelling evidence that student ratings and student learning and teaching effectiveness are highly correlated. The reason may in part be due to the fact that evidence suggests that student ratings of their instructors are subject to a range of potential biases. A major concern is questions about the statistical validity of SETs, and it can be misleading to put too much weight on averages, particularly combined averages, as ratings of teaching effectiveness, or fractional differences in scores that can fall within expected variation (N/A, 2022). Research identified a number of factors that affect SETs that are out

of instructors' control, which renders the results/ratings awfully invalid, unreliable, and biased. These factors include, but are not limited to, variables such as an instructor's gender, perceived attractiveness, ethnicity/race, etc. (Boring, 2015; N/A #1, 2022).

### **SETs vs an Instructor's Gender**

Research has increasingly questioned the validity, reliability, and impact of SETs as an indicator of teachers' performance and assessment (Okoye et al, 2020; Boex, 2000; Boring, 2017; Drowling, 2000; Rivera and Tilesik, 2019). Also, Rivera et al (2019) observe that researchers have paid little attention to the architecture of assessment and how biased evaluations play a primary role in sustaining gender inequalities (Okoye et al, 2020). Most recent studies provide new evidence of gender bias in teaching evaluations (Boring, 2017; Okoye et al, 2020; Heath et al, 2019; Islahi and Nasreen, 2013; Kafedice, Maler, and Niksic, 2018; Laube, Massoni, Sprague, and Ferber, 2007; Mengel, Sauermann, and Zolitz, 2019; Rivera and Tilesik, 2019; Sabbe and Aelterman, 2007). In the above studies, there are several facts about the differences in teaching qualities and performances of male and female teachers (Okoye et al, 2020). Sabbe et al (2007) and Okoye et al (2020) declare that female and male teachers are often presumed to differ in teaching style, capacities, and effects considering the students' perspectives and teaching as a career.

“The existence of presumed incompetence that affect both women of color and white women should provide a basis for a deeper understanding, sisterhood and alliance among women and enable work across racial lines to combat the presumption as well as other professional issues. But women can only forge that bond by acknowledging – rather than ignoring – the differences in the presumption's operation.” (Armstrong and Wildman, 2014, p 220)

Many studies agree that females receive systematically lower teaching evaluations than their male counterparts (Haffernan, 2022; Merritt, 2008; Boring, 2017; Okoye et al, 2020) and declare that male teachers are perceived by both male and female students as being more knowledgeable and having more persuasive class leadership skills, even though the students appear to learn more from female than from male teachers. These assumptions are reflected in the SETs. In addition, male students in particular discriminate in favor of male professors (Boring, 2017). The extract below is from a student evaluation of a female teacher:

“I want you personally to know I have hated every day in your course and if I wasn't forced to take this, I never would have. Anytime you mention this course to anyone who has ever taken it they automatically know that you are a horrific teacher, and that they will hate every day in your class. Be a human being shows some sympathy everyone hate this class and the material so be realistic and work with people.” Excerpt from a student e-mail to a female online professor. (Mitchell and Martin, 2018, p 1)

“The data are clear: a man received higher evaluations (SET[s]) in identical courses, even for questions unrelated to the individual instructor's ability, demeanor, or attitude” (Mitchell and Martin, 2018:4). “... in their current form, SET's might constitute another form of 'Weephole' for women and minorities in academic career pipelines that structurally contributes to higher attrition and lower achievement” (Chavez et al, 2019, p 4). “Frankly, as a sociologist specializing in gender, we are puzzled by conclusions that gender has no impact on teaching evaluation. Three decades of scholarship have found gender to be a significant factor in shaping interactions, practices, and outcomes in every major realm of human social life – why would the classroom be any different? (Laube et al, 2007, p 90).

While white women are members of an oppressed group based on gender, they still experience privilege based on race. This dual oppressor/oppressed identity often becomes a root of tension when white women are challenged to consider their white privilege by women of color (Accapadi, 2007, p 208). For women of color, “I noted the dual discrimination I felt as a black female. To me the discrimination was multiplicative, not additive. In other words, I was black times a woman every day, not black plus a woman,

which implies you may be able to subtract my identity. The discrimination I felt was against me as holistic black woman” (Wing, 2014, p 219).

### **SETs vs Race/Ethnicity Bias**

Just as gendered evaluations operate on “shifting standards” where one making a judgment is compelled to do so relative to a reference point, we assert that the same process occurs with people of color and accent in comparison to white males with native linguistic inflections (Chavez and Mitchell, 2019, p 4). Students’ expectations of their professors can conflict with expectations they have of people of different ethnic groups (Radhakrishnan, Frederikson, and Toh, 2021). Stereotyped expectations can also explain why professors of ethnic minority groups are rated lower than white professors, with each ethnic group evoking different expectations based on the stereotype of that group (Radhakrishnam et al, 2021, p 4). Professors with East Asian last names are rated as less clear, having less English language proficiency, competent, cold (not sociable), and submissive (Subtirelu, 2015; Lin, Kwan, Cheung, and Fiske, 2005; Radhakrishnan et al, 2021). When Asian professors violate the stereotype of the ideal Asian by being “dominant” and “warm,” they are rated as less helpful and lower on interpersonal skills than white and African American professors (Reid, 2010; Subtirclu, 2015; Bavish, Madera, and Hebl, 2010; Sidanius et al, 1989; Radhakrishnan et al, 2021).

Racial biases expressed in SETs reflect historical stereotypes, but the lens of racial microaggressions provides a more contemporary framework for understanding (Caldwell, 2018). Racial microaggressions are defined as a series of mini-assaults and are a challenging aspect of prejudice to clearly identify due to their often indirect, subtle, or unintentional nature (Feagin and McKinney, 2003; Smith, Allen, and Danley, 2007; Caldwell, 2018). They reflect much of the changing displays of racism since the civil rights movement where more overt behaviors were replaced with covert or aversive presentations (Dovidis, Gaertner, Kawakami, and Itodson, 2002). Professors of color face “a predisposition to assess minority performance in a negative or hypercritical fashion, an intolerance for even small mistakes committed by minorities, [and] a proclivity towards denying minorities the difference or presumption of competence accorded to White males – anti-minority mindset – is a set of negative biases held against minorities either consciously or unconsciously” (Brooks, 1987, pp 2–3).

The presumption of incompetence survived the displacement of black teachers under desegregation and still impacts black teachers and professors today (Smith, 1999, p 96). These assumptions are baseless where competence to teach is concerned. Just for the record, do not assume that someone will be your confidant just because he or she tend to be “friendly” talking to you or are the same race/gender as you. I heard a true story where a well-published professor of color went up for promotion to full professor rank and only two out of 21 voting members of the promotion committee who voted against his promotion were a person of color and a white person whom the faculty of color thought was his friend. There are people who resent a rising star and/or a person of color being their equal.

### **SETs vs Tools of Oppression**

According to Wang and Gonzalez (2020), quantitative studies show that white American professors receive significantly higher SET ratings than non-white professors, and they describe the non-whites to include African Americans, Asian Americans, Hispanic Americans and people of foreign origin, which implies the presence of bias in SETs. Smith and Hawkins (2011) agreed and revealed other studies involving exploration of the effect of race and ethnicity bias on the outcomes of SETs by comparing SET ratings of Hispanic, African American, and Asian faculty to white faculty – for universities to recruit and maintain marginalized faculty, they must respond to biases (race, gender, etc.) surrounding SETs (Smith, 2007; Hamermesh and Parker, 2005; Anderson and Smith, 2005).

SETs enable forms of interpersonal violence that appear neutral because they are based on numbers and data. The same racial epistemologies, gendered expectations, and other biases that inform student responses in SETs also inform how individual faculty and administration interpret SETs and actively translate them into cases for or against contract renewal and tenure (Rodriguez 2019, p 3). According to Caron (2018),

“Student evaluation become weapon. I have seen them used to try to kill the chances of tenure for a Hispanic female professor who had already won a teaching award. I have seen them ignored for White male professors who have strong publications. I have seen them used to kill faculty candidates. Worse, I hear stories that mirror the ones I have observed. In short, I have seen student evaluations used by associate deans and students alike the way a chimpanzee might use a lightbulb: Not for illumination!” (p i)

The biases in student evaluation are used to reinforce the biases of the department seeking to deny tenure to professors who aren't white males (Caron, 2018). SETs are perfect vehicles for a type of race/gender-blind discrimination because they allow one to claim detachment and objectivity. They pretend the “best qualified” are measured and confirmed through a natural process that just so happens to confirm the worst stereotypes about minorities and women (Ray, 2018).

Using biased evaluations allows colleges and universities to punish those whose identities deviate from white male normativity (Ray, 2018). Why did the University of Southern California (USC) eliminate student evaluations in faculty promotion and tenure decisions? You be the judge. Students will still evaluate their professors with some adjustments, but those ratings will not be used in high stakes personnel decisions at the USC (Caron, 2018). Administrators and faculty peers don't nearly care enough to spend the time reading or assessing written evaluations, but prefer the “pointless and stupid” multiple choice types, which allow the obtaining of numerical scores with little effort or meaning. Of course, this is why every university is administering SETs: they don't really care about quality; they just want an easy number to get, and, among other things, this facilitates getting rid of minorities and/or “troublemakers” if they're looking for some excuse – which they are (Lawrence, 2021, p 5). SETs are being misused against demanding or unpopular faculty members (Francisco, in Rodriguez, 2019). The misuse of SETs is not limited to tenure and promotion decisions; SETs are also misused in post-tenure evaluations. Indeed, SETs are so unreliable that their use is harmful to universities, which, as institutions partly devoted to research, should learn from research not to use SETs in personnel action decisions (Boxer, in Rodriguez, 2019). “Student evaluations of teaching (SETs) can be weaponized to justify undermining academic freedom and subjecting (minority) interviewed and contingent faculty to surveillance and bullying” (Rodriguez 2019, p 1). “... we don't even know what the hell we are measuring with student evaluations ... maybe it is just time to take away the weapon” (Caron, 2018, p 1).

Many studies that we have researched for this paper criticize SETs as biased or a poor measure of teaching effectiveness, or both. Our research has cut through contradictory studies of SETs and reflected/touched on reliability, validity biases, consumerism, and tools of oppression, and has made some recommendations for future changes.

## DISCUSSION AND CONCLUSION

Using SETs to determine how to evaluate, promote, tenure, and pay professors negatively affects ethnic minority and women professors' hiring, retention, representation, and advancement in academia. Because many university administrators tend to use SET rates as a primary mechanism to assess faculty teaching, and since these assessments are tied to compensation, academic entitlement can entice faculty to lower academic standards. Administrators in higher education must ensure meaningful learning is best served by maintaining academic standards even where there is some lessening of student and teacher comfort levels (Stewart, 2009, p III) or it goes against SET rates.

For all faculty of color, I do not think we can often change how other people label us (as presumed incompetent); that is their problem, even though their problem may cause us pain (Reyes, Kupenda, Onwuachi-Willig, Wildman, and Wing, 2014). Accepting that we may never be truly valued at our institutions may be discouraging. Therefore, after tenure, we must challenge ourselves to continue to grow, mentor next generations, and radiate joy within and outside our institutions (Reyes et al, 2014, p 244).

We want to be clear. We do not believe assessment is a bad thing, as we need it in the teaching and learning process, but when there are multiple methodological flaws, invalid results, products of cultural and other biases, and the weaponizing of results/rates against minority faculty members, we may just need to get rid of the weapon until we can figure out the appropriate way to measure teaching and learning effectiveness.

## RECOMMENDATIONS

- This study would like to recommend that students should not or cannot rate professors' teaching but may provide useful feedback on their perceptions and experiences.
- Considering the impact associated with students' customer mentality of entitlement and expectations in the classroom, higher education administrators should take a closer look at these influences on SETs and not use them for personnel action decisions.
- Administrators should never rely on these evaluations as a sole measure of teaching effectiveness. Peer evaluation of teaching, comprehensive teaching portfolios, and syllabus and other teaching material could complement SETs.
- SETs penalize faculty who are minorities, hardworking, demanding, and/or women; hence, a good method of evaluation needs to be developed, like setting up a committee of experts to review/evaluate faculty members' teaching effectiveness.
- Using SETs to determine how to evaluate, promote, and pay negatively affects minorities and women professors' retention and advancement in academia; hence, universities should remove SETs from personnel decisions in order to reduce systematic institutionalized discrimination.
- A good SET should be figured out that is well designed and very low in bias to analyze data carefully, taking note of patterns every period, checking for proven bias, and using multiple evaluation sources of data, including, but not limited to, administrative evaluations, peer evaluations, self-evaluations, and course material evaluations.
- Considering the impact and peripheral factors associated with students' consumer mentality, entitlement, and expectations in the classroom, it is time for administrators in higher education to take a closer look at these influences on SETs, examine what is really going on when SETs are done, and respond accordingly (Burke et al, 2019, p 39).
- Higher education should redefine customer service and craft policy with the public not the student most in mind (Searcy, 2017).
- It is common knowledge to educators and administrators that faculty members of color have to navigate the "inequitable terrain of academia," including confronting students' stereotyping of them and evaluation systems that disadvantage them. Hence, tangible support for faculty of color should be offered, including greater opportunities for mentorship and networking, and emphasis on connecting with other faculty of color and allies (Martinez, Chang, and Welton, 2017, p 708; Ortega, Leston, and Rodriguez-Soto, 2014; Radhakrishnan, 2021).
- The COVID-19 pandemic has dramatically altered teaching methods and may negatively affect SETs of some faculty members, although some younger faculty may have found the transition to online teaching to be easier than their more senior colleagues. Administrators should be particularly cautious in how SETs are incorporated into personnel decisions during this period (Kreitzer and Sweet-Cushman, 2022).
- Minorities and women faculty should speak up and stop abuse and torture (Reye, 2014).

## REFERENCES

- Allen, W. R., Epps, E. G., Guillory, E. A., Suh, S. A., & Bonous-Hammarth, M. (2000). The Black Academic: Faculty Status Among African Americans in U.S. Higher Education. *Journal of Negro Education*, 69(1-2), 112–127. Retrieved from <https://psycnet.apa.org/record/2002-10183-007>
- Accapadi, M.M. (2007) When White Women Cry: How White Women’s Tears Oppress Women of Color, *The College Student Affair Journal* 26 (2) p 208-215 (Google Scholar)
- Armstrong and Wildmen (2014) Working across Racial Lines in a Not-So-Post- Racial World, In Maritza Ryes, (Ed) Presumed Incompetent – The Intersections of Race and Class for Race and Class for Women in Academia Symposium, Plenary Panel, Retrieved from [tps://commons.law.famu.edu/cgi/viewcontent.cgi?article=1122&context=faculty-research](https://commons.law.famu.edu/cgi/viewcontent.cgi?article=1122&context=faculty-research)
- Abel, R.L. (1990) Evaluating Evaluations: How should Law Schools Judge Teaching, 40J. Legal EDUC. 407-437. *HEINOLIN* Retrieved from <https://heinonline.org/HOL/LandingPage?handle=hein.journals/jled40&div=40&id=&page=>
- Aleamoni, L.M (1999) Student Rating Myths Versus Research Facts from 1924 to 1998 *Journal of Personnel Evaluation in Education* 13 no 2 p 153-166
- Anderson K.J. and Smith, G. (2005) Students’ Preconceptions of Professors: Benefit and Barriers According to Ethnicity and Gender, *Hispanic Journal Behavioral Science*, Vol. 27, issue 2.
- Boring, A (2017) Gender biases in student evaluations of teaching, *Journal of Public Economics* 145, 27-41 Retrieval from <http://www.sciencedirect.com/science/article/pii/S0047272716301591>  
<https://www.sciencedirect.com/science/article/abs/pii/S0047272716301591>
- Boring, A., (2015). Gender Bias in Student Evaluations of Teachers, *OFCE-PRESAGE- Sciences-Po Working Paper*, Retrieved from <http://www.ofce.sciences-po.fr/pdf/dtravail/WP2015-13.pdf>
- Baker, P. and Copp, M., (1997) Gender Matters Most : The Interaction od Gendered Expectations, Feminist Course Content, And Pregnancy in Student Course Evaluations, *Teaching Sociology* , Vol 25, 29-43., Retrieved from <https://www.jstor.org/stable/1319109?seq=1>
- Benton, S.L.D, Dacon, D., and Pallett, W.H., (2013) Validity of Student Self- Reported Ratings of Learning “Assessment & Evaluation in Higher Education 3894 p. 377-388
- Beam, K. S. (1989) The Gender Gap in Law School Classroom Beyond Survival, *14 VT REV* 23 29
- Borch, I., Sandvoilr and Risor, (2020) Discrepancies in Purposes of Student Course evaluations: What does it mean to be “satisfied”. *Educational Assessment Evaluation and Accountability* 3283-102 Retrieved from <https://eric.ed.gov/?id=EJ1247575>
- Bavistic, A., Madera, J.M., and Hebl, M.R. (2010) The Effect of Professor Ethnicity and Gender on Students Evaluations: Judged Before Met, *Journal of Diversity in Higher Education* 3, no 4: p 245-256
- Brooks, R.L., (1987) Anti-Minority Mindset in the Law School Personnel Process Towards an Understanding of Racial Mindset, 5 (1) *Law & Inequality* 1 Retrieved from <https://scholarship.law.umn.edu/lawineq/vol5/iss1/1/>
- Boex, L. (2000) Attributes of Effective Economics Instructors: An Analysis of Student Evaluations. *The Journal of Economic Education*, Vol. 32, No. 3, 211-227
- Burke, M.C., Hughey, A. W. and Cannonier, C., (2019) The Impact of Consumer and Academic Entitlement Attitudes on Student Course and Faculty Evaluations: A Call for Change to Academic Leaders, *International Journal of Leadership and Change Vol 7 Issue I*
- Czekanski, K., Hoerst, B. J., & Kurz, J. (2018). Instituting evidence-based changes to improve first-time NCLEX-RN® pass rates. *Journal of Nursing Regulation*, 9(1), 11-18. Retrieved from: [https://doi.org/10.1016/S2155-8256\(18\)30049-8](https://doi.org/10.1016/S2155-8256(18)30049-8)
- Chowning, K., and Campbell, N. J. (2009) Development and Validation of a Measure of Academic Entitlement: Individual Differences in Students' Externalized Responsibility and Entitled Expectations. *Journal of Educational Psychology*, 101 p 982-997
- Caron, P., (2018) Weaponizing Student Evaluation, *Tax Prof Blog Blue Journal* Retrieved from [https://taxprof.typepad.com/taxprof\\_blog/2018/09/weaponizing-student-evaluations.html](https://taxprof.typepad.com/taxprof_blog/2018/09/weaponizing-student-evaluations.html)
- Clayson, D., (2009) Student Evaluations of Teaching: Are They Related to What Students Learn? *Journal of Marketing Education* 31(1) p16-30
- Cox, S., Rickard, M. K., and Lowery, C., (2021) The Student Evaluation of Teaching: Let’s be Honest – Who is Telling the Truth? *Marketing Education Review*. Vol. 32, Issue 1.
- Caldwell, K.J., (2018) Racial Bias and Student Ratings of Instructors, TEACHING SURVEY, Pitt: University Center for Teaching and Learning Retrieved from [https://teaching.pitt.edu/wp-content/uploads/2018/12/OMET-Racial\\_bias\\_and\\_student\\_ratings\\_of\\_instructors.pdf](https://teaching.pitt.edu/wp-content/uploads/2018/12/OMET-Racial_bias_and_student_ratings_of_instructors.pdf)
- Clayson, D. E., (2021), *A Comprehensive Critique of Student Evaluation of Teaching: Critical Perspective on Validity, Reliability and Impartiality*, Routledge, Taylor & Francis Group, eBook.
- Contra, J. (1993) *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness*, (Jossey Bass higher and adult education series), San Francisco: Jossey Bass Inc Pub.
- Chavez, K., and Mitchell M.W. K (2019) *Exploring Bias in Student Evaluations: Gender, Race, and Ethnicity*, Cambridge University Press.
- Drowling, W. (2000) Why We Should Abolish Teaching Evaluations. *The Daily Targum*, Retrieved from [www.wcdrtugers.net/targeval.htm](http://www.wcdrtugers.net/targeval.htm)
- Denzin, N., and Lincoln, Y. (2011) *The Sage Handbook of Qualitative Research Association*, Retrieved from: [https://scholar.google.com/scholar?q=Denzin,+N.,+and+Lincoln,+Y+\(2011\)+The+sage+handbook+of+qualitative+research+Association&hl=en&as\\_sdt=0&as\\_vis=1&oi=scholart](https://scholar.google.com/scholar?q=Denzin,+N.,+and+Lincoln,+Y+(2011)+The+sage+handbook+of+qualitative+research+Association&hl=en&as_sdt=0&as_vis=1&oi=scholart)

- Denman, C. L., & Cohn, T. M. (2022). Use of Standardized Testing to Predict NCLEX-RN Success for Associate Degree Nursing Students in a Concept-Based Curriculum. *Teaching and Learning in Nursing*, 17(4), 378-382. Retrieved from: <https://doi.org/10.1016/j.teln.2022.05.001>
- Dabler-Larsen, P. (2009) Learning Oriented Educational Evaluation in Contemporary Society. In Ryan K. E.,(Ed), *The Sage International Handbook of Educational Evaluations* 307-322 Los Angeles, (Calif)
- Dano, T. and Stensaker, B., (2007) Still Balancing Improvement and Accountability! Development in External Quality Assurance in the Nordic Countries 1996-2006. *Quality in Higher Education* 13(1) 88-93.
- Dovidio, J.F., Gaertner, S.E., Kawakami, K. and Hodson, G (2002) Why Can't We Just Get Along? Interpersonal Biases and Interracial Distrust. *Cultural Diversity and Ethnic Minority Psychology Journal* 8(2) Retrieved from <https://psycnet.apa.org/record/2002-02415-002>
- Edmunson, M. (1997) On the Use of Liberal Education as Lite Entitlement for Bored College Students, *Harper's Magazine*, 295. p 39-49. Retrieved from <https://www.ljhammond.com/essay.htm>
- Esarey, J and Valdes, N (2020) Unbiased, Reliable and Valid Student evaluations Can Still Be Unfair. *Assessment & Evaluation in Higher Education*, Vol 45, Issue 8, Retrieved from <https://www.tandfonline.com/doi/full/10.1080/02602938.2020.1724875>
- Feagin, J.R., and McKinney, K.D. (2003) *The Many Cost of Racism*, Rowman & Littlefield, Lanham, Md.
- Face, G.M.F (June 23, 2019) The Customer is Always Right is Awfully Wrong. *The Canberra Times*, Retrieved from <https://www.canberratimes.com.au/story/6219583/heres-why-the-customer-is-always-right-is-awfully-wrong/>
- Freishtat, R. L., (2016) Expert Report on Student Evaluations of Teaching, *Ryerson Faculty Association and the Ontario Confederation of University Faculty*, Retrieval from [https://ocufa.on.ca/assets/RFA.v.Ryerson\\_Freishtat.Expert.Supplemental.Reports\\_2016.2018.pdf](https://ocufa.on.ca/assets/RFA.v.Ryerson_Freishtat.Expert.Supplemental.Reports_2016.2018.pdf)
- Fook, J., Ryan, M. and Hawking, L. (2000) *Professional Expertise: Practice, Theory and Education for working in Uncertainty*, London: Writing and Birch
- Frazier K.N. (2011) Academic Bullying: A Barrier to Tenure and Promotion for African American Faculty. *Florida Journal of Educational Administration Policy*, Fall Vol 5 Issue 1, p. 1-13
- Galbaith, C., Merrill, G., and Kline, D. (2012) Are Student Evaluations of Teaching Effectiveness Valid for Measuring Student Learning Outcomes in Business Related Classes? A Neutral Network and Bayesian Analyses. *Research in Higher Education*, Vol. 53(3) 353-374
- Hobson, S. M. and Donna, M.T. (2001) Understanding Student Evaluations: What All Faculty Should Know, *College Teaching*, Vol. 49, no 1. p 26-31
- Heath J.K., Weissman, G.E., Clancy, C.B., Shou, H., Farrar, J.T., & Dine, C.J. (2019) Assessment of Gender-Based Linguistic Differences in Physician Trainee Evaluations of Medical Faculty Using Automated Text Mining. *JAMA Network Open*, 2(5)
- Higgs, K., (2022) A Brief History of Consumer Culture, *The MIT Press*.
- Hefferman T. (2022) Abusive Comments in Student Evaluations of Courses and Teaching: The Attacks Women and Marginalized Academics Endure, *High Education*, Retrieved from <https://link.springer.com/article/10.1007/s10734-022-00831-x>
- Hamermesh D.S. and Parker, A (2005) Beauty in the Classroom Instructor's Pulchritude and Putative Pedagogical Productivity, *Economics of Education Review* 24 (4) 369- 376.
- Himmeibach, V, (March 19,2019) Course Evaluation: What's Wrong and How to Fix it., Retrieved from <https://tophat.com/blog/course-evaluation/>
- Hollis L D (2022) In the Room, But No Seat at the Table: Mixed Methods Analysis of HBCU Women Faculty and Workplace Bullying. *Journal of education*, Retrieved from <https://journals.sagepub.com/doi/abs/10.1177/00220574221102329>
- Huemer, M., (2001) Student Evaluation a Critical Review. *PhilPaper*, Retrieved from <http://www.owl232.net/papers/sef.htm>
- Islahi, F., and Nasreen, N. (2013) Who Make Effective Teachers, Men or Women? An Indian Perspective, *Universal Journal of Educational Research*, 1(4) p 285-293
- Johnson, V.E., (2003) *Grade Inflation: A Crisis in College Education*, New York, Springer
- Jan, T (April 7, 2010) Getting Students to Make Their Mark: Fewer Bothing to grade Professors as System Go Online. *The Boston Globe, Boston.com*.
- Kupenda in Reyes, M., (2014) Reflections on Presumed Incompetent The Intersections of Race and Class for Women in Academia. Symposium. *The Plenary Panel, Scholarly Commons @ FAMU* from <https://link.springer.com/article/10.1023/A:1026447822732>
- Kopinak, J., (1999) The Use of Triangulation in a Study of Refugee Well Being, *Quality and Quantity, International Journal of Methodology Springer*, Vol. 33 (2) 169-183.
- Kafedric, E., Malec, D., and Niksic, E., (2018) Differences Between Male and Female Secondary School Students in Assessing Their Physical and Health Education Teachers' Competence, *Sport Science* 11(1) p. 64-70
- Kreitzer, R. J., and Sweet-Cushman, J., (2022) Evaluating Student Evaluation of Teaching: A Review of Measurement and Equity Bias in SETS and Recommendations for Ethical Reform, *Journal of Academic Ethics* 20 (1) p. 73-84
- Lippmann, S., Bulanda, R.E, and Wagenaar, T. C. (2009) Student, Entitlement Question Issues and Strategies for Confronting Entitlement in the Classroom and Beyond. *College Teaching*, 57 p. 197-204.
- Li, D; Benton, S.L., Brown, R., Sullivan, P., and Ryalls, K. R;(2016) Analysis of idea Student Ratings of Instruction system 2015 Pilot Data. IDEA Technical Report No19 Assessed 1 October 2019, *ERIC*, Retrieved from <https://eric.ed.gov/?id=ED573674>

- Lim, M. H., Kwan, V.S.Y., Cheung, A., and Fiske, S.T. (2005) Stereotype Content Model Explains Prejudice for an Envied Outgroup: *Scale Psychology Bulletin* 31, no. 1, p 34-47
- Lawrence, J. W. (2021) Student Evaluations of teaching are not valid. *American Association of University Professors*. Retrieved from [https://www.aaup.org/article/student-evaluations-teaching-are-not-valid#.Y\\_zgcXbMK70](https://www.aaup.org/article/student-evaluations-teaching-are-not-valid#.Y_zgcXbMK70)
- Laube, H., Massoni, K., Sprague, J., and Ferber, A. (2007) The Impact of Gender on the Evaluation of Teaching: What we know and what we can do. *NWSA Journal* 19(3), p 87-104
- M.A Martinez., Chang, A., and Welton, A.B. (2017) Assistant Professor of Color Confront the Inequitable Terrain of Academia? A Community Cultural Wealth Perspective, *Race, Ethnicity and Education* 20(5) p 696-710
- March, H. W., (1987) Student Evaluations of University Teaching: Research Findings, Methodological Issues and Directions for future research, *International Journal of Educational Research*, 11, 253-369.
- Mason A. D., (2010) Police Culture and Ethics Toward an Understanding and Expansion of Police Culture and Ethical Research, *Graduate Theses and Dissertation paper* 11482 Retrieved from <https://dr.lib.iastate.edu/entities/publication/242a0143-5814-4e98-95e2-1a1df074de5d>
- Merritt, D.J (2008) Bias, the Brain and Student Evaluations of Teaching, *John, S's Law Review*, Volume 82 #1
- Mengel, F., Sauermann, J., and Zolitz, U., (2019) Gender Bias in Teaching Evaluations, *Journal of the European Economic Association*, 17(2) p 535-566
- Mitchell, M.W., and Martin, J., (2018) Gender Bias in Student Evaluations: *Political Science and Politics* 51 (3)1-5
- Norris, D. (2019). Assessing the Learned Learner When Using a Concept Curriculum in Nursing Education. *Research Issues in Contemporary Education*, 4(n1) 35-44
- N/A #1 (2022) Course Feedback as a Measure of Teaching Effectiveness, *Stanford Evaluation & Research*, Retrieved from <https://evals.stanford.edu/evaluating-teaching/course-feedback-measure-teaching-effectiveness>
- Ortega - Luston, R., & Rodriguez - Soto, I. (2014) Challenges, Choices and Decisions of Women in Higher Education: A Discourse on the future of Hispanic, Black, and Asian Members of the Professoriate, *Journal of Hispanic Higher Education*, 13 (4) p285-302
- Onuehergie, A., Daniel, E., and Collins, K. (2009) A Meta-Validation Model for Assessing the Score-Validity of Student-Teaching Evaluations. *Quality and Quantity*, 43(2), p. 197-209
- Olabi, R. M., (2021) The Pregnancy Drop: How Teaching Evaluations Penalize Pregnant Faculty, Humanities and Social Sciences Communications, Article number 253.
- Okoye, K., Arrona-Palarios, A., Camacho-Zuniga, C., Hammout N., Luttmann, E., Nakamura, J. E., and Hosseini, S., (2020) Impact of Students Evaluation of Teaching: A Text Analysis of the Teachers Qualities of Gender, *International Journal of Educational Technology in Higher Education* 17(49).
- Pounder, J. S., (2007) Is Student Evaluation of Teaching Worthwhile? An Analytical Framework for Answering the Question. *Quality Assessment in Education* Vol 15 No 2.
- Patitu, C., and Henton, K., (2003) The Experience of African American Women Faculty and Administrators in Higher Education: Has Anything Changed? *New Direction for Student Services*, 104. *Wiley Periodical Online*.
- Pope, D. C. (2001). "Doing School": *How We are Creating a Generation of Stressed Out, Materialistic, and Miseducated Students*. New Haven, CT: Yale University Press
- Rodriguez, J., (2019) The Weaponization of Student Evaluations of Teaching: Bully and the Undermining of Academic Freedom, *Journal of Academic Freedom*, Vol 10
- Raban, C., (2007) Assurance "Versus" Enhancement: Less is More *Journal of Further and Higher Education* 31(1) 77-85. Retrieved from <https://www.semanticscholar.org/paper/Assurance-versus-enhancement%3A-less-is-more-Raban/0ceb8a92d16aa073633b5aca4314227916a3a098>
- Rempel, R. J., (2018) The Validity and Utility of Student Evaluation, *Perspective on Social Work*, Volume 13 no.1
- Reid, I.D., (2010) The Role of Perceived Race and Gender in the Evaluation of College Teaching on RateMyProfessor.com *Journal of Diversity in Higher Education* 3, no 3. P 137-274
- Reyes, M., Kupenda, A. M., Onwuachi-Willig, A., Wildman, S. M., and Wing, A. K, (2014) Reflection on Presumed Incompetent: The Intersection of Race and Class for Women in Academia Symposium – The Plenary Panel, Scholarly Commons @ FAMU Law. Retrieved from <http://digitalcommons.law.scu.edu/facpubs>
- Radhakristnan, P., Frederickson, M., and Toh, S.M (2021) How Student Evaluations of teaching are Biased Against Women and Ethnic Minority Professors. *ResearchGate*. Retrieved from [https://www.researchgate.net/publication/348152682\\_How\\_Student\\_Evaluations\\_of\\_Teaching\\_are\\_Biased\\_against\\_Women\\_and\\_Ethnic\\_Minority\\_Professors](https://www.researchgate.net/publication/348152682_How_Student_Evaluations_of_Teaching_are_Biased_against_Women_and_Ethnic_Minority_Professors)
- Ray, V (February 9, 2018) Is Gender Bias an Intended of Feature of Teaching Evaluation? *Inside Higher ED*.
- Rivera, L.A. and Tilesik, A. (2019) Scaling Down Inequality: Rating Scales, Gender Bias and the Architecture of Evaluation. *American Sociological Review* 84 (2) p 248-274
- Remmars, H. H. and Brandenburg, C. C. (1927) Experimental data on the Purdue Rating Scale for Instruction. *Educational Administration and Supervision*, 13, 519-527
- Remmer, H. H. (1933) Learning, Effort, and Attitudes as Affected by three methods of Instruction in Elementary Psychology, *Purdue University Studies in Higher Education (Monograph No 21)*
- Shulman, L (1988) Disciplines of Inquiry in Education: An Overview. In R. M. Jaeger (Ed) *Complimentary Methods for Research in Education* (p3-7) *American Educational Research Association*, Retrieved from <https://www.aera.net/Publications/Books/Complementary-Methods-for-Research-in-Education>



- Spooren, P., Brock, B. and Mortelman, D.;(2013) On the Validity of Students Evaluation of Teaching: The state of the Art. *Reviews of Educational Research*, 83(4) p.598-642
- Seldin, P., (1984) *Changing Practices in Faculty Evaluation*, Jossey Bars, San Francisco, CA
- Seldin, P (1993) The Use and Abuse of Student Rating of Professors, *The Chronicle of Higher Education*, Vol 39 No 46. A40
- Searcy Y. D. (June 5, 2017) Let Me Speak to the Manager: Higher Education Should View the Public Not the Student as the Customer, *INSIDE HIGHER EDUCATION*
- Smith P.J. (1999) Teaching the Retrenchment Generation: When Sapphire Meet Socrates at the Intersection of Race, Gender and Authority, 6 WM + Mary J. Women + L53, 162-163.
- Sieverding M, Eib. C., Neubauer A. B. + Stahl T., (2018) Can lifestyle preferences help explain the persistent gender gap in academia? The “mother work less” hypothesis supported for Germany but not for US early career researchers: PLOS ONE 13 (B): e0202728.
- Swann J.M. (2022) Academic Bullying and Diversity: Challenges and Solutions, *The Scientist Forum, FEBS Letters/Volume 596*, Issue 221p 2855-2858
- Spooren, P. Brockx, B., and Mortelmans, D. (2013), On the Validity of Student Evaluation of Teaching: The State of the Art. *Review of Educational Research*, 83(4), 598-642
- Sloman, S.A., (2002) Two Systems of Reasoning: in Gilovich, T and Kahneman, D., (Ed) *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press 279
- Stroebe, W. (2020) Student Evaluation of Teaching Encourages Poor Teaching and Contribute to Grade Inflation: A theoretical and Empirical Analysis, *Basic and Applied Social Psychology Vol. 42 Issue 4*
- Stroebe, W. (2016) Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluation, *Perspectives on Psychological Science*, Vol 11(6) 800-816
- Sidanius, J and Crane, M (1989) Job Evaluation and Gender: The Case of University Faculty: *Journal of Applied Social Psychology* 19, no 2 p 174-197
- Subtirelu, N.C. (2015) She Does Have an Accent but – Race and Language Ideology in Student’s Evaluations of Mathematics Instructors on RateMyProfessors.com. *Language in Society* 44 no 1. p 35-62
- Smith, B.P. (2007) Student Ratings of Teaching Effectiveness: An Analysis of End of Course Faculty Evaluation, *College Student Journal*, 41, p 788-800
- Searcy Y.D. (2017) Let me speak to the manager! Higher Ed should redefine customer service, and craft policy with the public not the student most in mind, *Inside Higher Ed Careers*, Retrieved from <https://www.insidehighered.com/views/2017/06/05/higher-ed-should-view-public-not-student-customer-essay>
- Smith, W.A., Allen, W.R. and Danley, L.L., (2007) “Assume the position --- You fit the description, *American Behavioral Scientist* 51(4).
- Sabbe, E., + Aelterman, A (2007) Gender in Teaching: A literature Services. *Teachers and Teaching*, 13(5) p 521-538
- Stewart, K (2009) Lessons For Teaching Millennials, *College Teaching* 57p 111-118
- Theall M., and Franklin, J., (2000) Creating Responsive Student Rating System to Improve Evaluation Practice, *New Directions for teaching and Learning* 83, 95-107
- Tillman, L.C., (2001) Mentoring African American Faculty in Predominantly White Institutions, *Research in Higher Education*, 42(3) p295-325.
- Trower, C. A., and Chait, R. P., (2002) Faculty Diversity: Too Little for Too Long, *Harvard Magazine*, Retrieved from <https://www.harvardmagazine.com/2002/03/faculty-diversity.html>
- Titus, J. (2008) Student Ratings in Consumerist Academy: Leveraging Pedagogical Control and Authority. *Sociological Perspectives* 5/(2): 397-422.
- Uttl, B, and Smibert, D (May 9, 2017) Student Evaluation of Teaching: Teaching Quantitative Courses Can be Hazardous to One’s Career. *Peer Journal*, e3299. Retrieved from <https://pubmed.ncbi.nlm.nih.gov/28503380/>
- Williams, J., (2016) Quality Assurance and Quality Enhancement: is there a Relationship? *Quality in Higher Education* 22(2) 97-102.
- Ward D (2021) An overview of research on use of student surveys to evaluate teaching. University of Kansas,
- Wing, A. K (October 29, 2014) Lesson from a Portrait: Keep Calm and Carry On, in Reyes, Maritza I., Angela Mae Kupenda, Angela Onwuachi-Willig, Stephanie M. Wildman, Adrien Katherine Wing, Eds. Reflections on Presumed Incompetent: The Intersections of Race and Class for Women in Academia Symposium - The Plenary Panel, 29 *Berkeley J. Gender L. & Just.* 195 (2014)
- Wang, L and Gonzalez, J. A. (2020) Racial/Ethnic and National Origin Biases in SET. *International Journal of Organizational Analysis*. Vol 28 No 4

