CHAPTER 7.

TEACHING TO FAIL? THREE FEMALE FACULTY NARRATIVES ABOUT THE RACIAL AND GENDER INEQUALITIES OF SETS

Mary Lourdes Silva
Ithaca College

Josephine Walwema
University of Washington, Seattle

Suzie Null
University of California, Santa Barbara

What does it mean to fail in an inequitable culture framed by the values and metrics of student evaluations of teaching (SET)? Failure to perform well on SETs can result in some form of administrative action or sanction, such as job loss, lack of promotion, or denial of merit pay or tenure (Spooren et al., 2013; Wachetel, 1998). Not only do negative SETs have a professional and financial cost, but negative SETs can impact self-esteem, self-efficacy, and faculty morale (Boswell, 2016; Kowai-Bell et al., 2012; Wachetel, 1998). Both female faculty and female faculty of color, in particular, are disproportionately impacted by SETs. For instance, multiple studies show that students define teaching effectiveness based on gendered personality characteristics—e.g., women as nurturing and men as brilliant (Sprague & Massoni, 2005; Storage et al., 2016); female faculty are rated similarly to male faculty at the beginning of a course up until the point female faculty critique student work in the form of grades (Buser et al., 2022); and students evaluate *effective* female faculty more negatively (Boring et al., 2016) and penalize female faculty of color for both their race and gender (Baslow, 1995; Boring et al., 2016, 2017; Davison & Price, 2009; MacNell et al., 2015; Mengel et al., 2017; Pittman, 2010). As a result, female faculty and female faculty of color are compelled to play on an unequal playing field, which has multiple professional and psychological ripple effects. When administration or faculty frame negative SETs as a failure of teaching rather than as a byproduct of a racist/sexist system of evaluation, female

DOI: https://doi.org/10.37514/PER-B.2024.2494.2.07

faculty and female faculty of color must allocate additional time and resources to improve their SETs, thereby detracting from research projects, publishing, professional service and networking, and self-care.

The validity and reliability of SETs have increasingly been called into question as a measurement of teaching effectiveness (Beran & Rokosh, 2009; Boring et al., 2016; Galbraith et al., 2012; Shevlin & Banyard, 2000; Spooren et al., 2013; Uttl et al., 2017). In a seminal review of SET research, Wachtel (1998) reports research findings that confirm and challenge the validity and reliability of SETs, underscoring variables such as time delivery of SETs, class time, level of course, class size, course electivity, workload, subject area, anonymity of student raters, gender, race, age, instructor rank, and personality. With each variable, it is critical to contextualize the results. For instance, large lecture classes may correlate with higher SET scores if students have elected to take a popular course with a reputable professor, whereas a large required first-year course typically correlates with lower SET scores in comparison to smaller seminar courses. Fifteen years later, a metastudy by Annan et al. (2013) confirms Wachtel's prior findings and indicates how factors outside faculty members' control can affect SET ratings. These can include the time of day, type of room, course level, course workload, course type (required versus elective), student attendance, students' keeping up with assigned reading, and expected grade in the course. Miles and House (2015) found similar results regarding class size, class type (required versus elective), and course grade expectations. Student perceptions of SETs and their value in higher education may also influence SET scores (Spooren & Christiaens, 2017). The validity of SETs is tenuous, at best, when 30 percent of students admit to submitting false information on SETs based on their personal opinion of the instructor (Clayson & Haley, 2011). Problems of validity and reliability are exacerbated for many female faculty and female faculty of color, who have to navigate a system with inherent and ingrained biases, making experiences of institutionally perceived failure a commonplace narrative.

The reliability and validity of SETs is more pronounced for female faculty because students unconsciously evaluate their abilities and intelligence by applying different criteria. Rivera and Tilcsik's (2019) study of student ratings of instructors found that the female instructor was rated "as less brilliant than her otherwise identical male counterpart" (p. 20). Moreover, according to Storage et al. (2016), women and African American faculty were less likely to be rated or perceived as "brilliant" or a "genius" among student respondents. Relatedly, MacNell et al.'s (2015) empirical study of gender bias in an online asynchronous course with multiple sections found that students rated more harshly instructors they assumed to be female. Mitchell and Martin (2018) report similar findings in their study of two identical online courses, one taught by a female professor

and the second by a male professor, with all variables held constant (i.e., course assignments, course format, and lectures). In a content analysis of student comments, Mitchell and Martin found that women were more likely to be judged for their appearance and personality and to be referred to as a "teacher" instead of "professor." Even when the male professor, on average, awarded lower grades, he received higher SET scores in comparison to his female colleague.

Evaluations based on gender are as problematic as other non-teaching criteria, but they are especially so when paired with race and ethnicity (Robinson, 2018; Reid, 2010). In underrepresented fields, women and minorities are more likely to report bias where their legitimacy is questioned (Dancey & Gaetane, 2014). On teaching evaluations, Bradley and Holcomb-McCoy (2004) report that a few Black female educators experienced "unequal treatment" and "racism" (p. 266). A study of non-white faculty by Lindahl and Unger (2010) showed that students' qualitative responses of these instructors were cruel, negative, and disrespectful, including several that, according to the authors, "were inappropriate to reprint" (p. 4). This antipathy is brought on by a variety of factors, the foremost of which is students' first impressions of instructors when based on race (Littleford et al., 2010; Roseboro, 2021; Taylor, 2021). This impression persists throughout the course, with students using it as the baseline to judge their instructors' expertise, estimations of authority, grasp of course material, and teaching style. Roseboro (2021) stresses that for Black faculty members (but also other BIPOC faculty and minority faculty), "being forced to analyze and include those course evaluations in the promotion and tenure application may, in fact, re-traumatize" (p. 57 of 190, Kindle Edition). For Roseboro, a minority instructor "requires a self-justification of one's right to be. And that prescribed self-justification reinforces the idea that one does not belong" (p. 57 of 190, Kindle Edition).

Ongoing feelings of failure and inadequacy due to SETs have psychological effects that impact professors' self-efficacy and self-esteem. In many cases, SETs may be the only feedback that faculty receive, particularly if constructive peer feedback is less available. Boswell (2016) found that participants who received positive SETs reported greater self-efficacy and confidence as professors and greater rapport with students. Boswell predicted that those with higher SET scores would most likely invest more effort to engage with students, whereas those with low SET scores may not have the motivation, affect, or resources to improve their SET scores. Negative evaluations, however, can shape professors' self-concept as it relates to teacher-student rapport (Kowai-Bell et al., 2012). Indeed, some faculty link their personal identity with negative student feedback (Arthur, 2009). Beran and Rokosh (2009) quote one instructor who described the process of SETs as "humiliating and frustrating" (p. 506). Yao and Grady (2005) found that faculty often experienced anxiety and nervousness when receiving SETs. They write that

faculty reported concerns about instructors lowering standards to improve SET scores (see also Wachtel, 1998). In a separate study by Crumbley and Reichelt (2009), 53 percent of accounting professors knew someone who engaged in defensive measures such as grade inflation, grading leniency, reduction of standards, and reduction of coursework to improve SET scores. Crumbley and Reichelt state that there are monetary and administrative penalties for low SETs; however, there is no penalty for grade inflation, coursework reduction, and grading leniency, which often become professors' only option within a broken system. Moreover, in some cases, higher SETs are actually correlated with lower levels of student achievement (Galbraith et al., 2012) or with worse student performance in subsequent courses (Kornell & Hausman, 2016). When faculty reduce grading standards or coursework as a defensive measure to preserve their jobs or obtain promotion, student learning can be compromised.

In summary, female faculty, and particularly female faculty of color, often receive statistically significantly lower ratings (Baslow, 1995; Boring et al., 2016), or they often need to invest more time and energy to get the same evaluations that men do (Owen, 2019). Boring et al. (2016) write that "SET appear to measure student satisfaction and grade expectations more than they measure teaching effectiveness" (p. 10). This puts many female faculty in the untenable situation of having to decide whether to compromise their research and publications agendas (which are also required for tenure and promotion at teaching colleges) to improve their SETs or risk getting low SETs to pursue a competitive research agenda.

West-Puckett et al. (2023) challenge binary constructs of failure in which there is a defined outcome of success. Failure in this context is experienced as shameful "red marks" that we are forced to carry with us from semester to semester or from one college to the next. How do we resolve the conundrum of how and when to fail with SETs if it is not possible to succeed in all areas of academia: teaching, research, and service? Moreover, how can we benefit from failure when the system of assessment is designed to privilege certain bodies, behaviors, ideologies, and practices and marginalize or oppress others?

In the rest of this chapter, we share our personal accounts with SETs and the psychological, emotional, pedagogical, professional, personal, and health consequences of working in this culture of failure. In the first narrative, Walwema shares her experiences of shame and its heavy toll at the start of her academic career, where low SETs compelled her to withdraw from a job search. Silva writes about the shame and anxiety experienced while serving multiple leadership roles in faculty development while repeatedly receiving lower evaluations in comparison to her white male subordinates. Null describes how gendered expectations in her female-dominated field created an untenable workload and stress concerning promotion and how a critical set of SETs compelled her to stop

teaching a particular course. These narratives will begin a conversation about the psychological, professional, and pedagogical consequences of navigating a field that leaves intelligent, skilled, experienced female experts constantly negotiating the conundrums of "failing" in one or more areas of their professional lives.

WALWEMA, THE RELUCTANT ACADEMIC

I embody all the attributes of foreignness. I speak differently, my outlook is global, my sensibilities are multicultural. Inhabiting these attributes has made me understand and accept the complexities and nuances of others. The position from which I operate is that of a Black woman migrant living in the United States. Here, I am constantly challenging both students and faculty about their essentialist assumptions about my identity and what it might say about me and where I am from. Occupying such a position is obviously problematic. And it is not unique to me. Thus, I acknowledge that while all of us are somewhat at the mercy of our histories, not all of us have the agency to embody the attributes of who we are and how we are perceived. I own these racialized and gendered attributes in the knowledge that I am always grounded in multiple marginalized realities.

Like other academics, I evaluate my performance based on external, often measurable criteria in the areas of teaching, research, and service. In a given year, I can enumerate the papers I have published, committees I have served on, and the work I have done on those committees; I can detail the classes I have taught and to what degree they have been successful. I can write at length the ways I have retooled aspects of the classes I have taught based on random surveys and sometimes direct solicitation of feedback from students during the course of the term. Which is why it is astounding that even with the abundance of research showing the malign nature of SETs, when it comes to decisions of hiring, promotion, retention, and tenure, colleges and universities still assign them an outsized influence. Being on the receiving end of SETs, often mean-spirited, insulting, and denigrating, with most bordering on the *ad hominem*, has induced in me a sense of failure. When I first handed out SETs at the end of my first teaching semester as a graduate student, I assumed I'd get feedback on three areas: things that went well, things that I ought to improve upon, and things that I definitely should eliminate. From that standpoint, I looked forward to using SETs to course-correct and retool. After all, I reasoned, this is the kind of feedback I gave my professors as a college student. Was I wrong! The SETs I encountered were mostly personal attacks ascribed to my race and gender. They were littered with phrases such as "does not know as much as she thinks," "she seems to care too much about writing" (I am a writing professor), and "I did not learn anything new." I was caught flatfooted. And, in the aftermath, I felt not

only disengaged, as though the subject of these SETs were someone other than me but also disempowered. I felt like I had lost my agency. What is it that empowers students to act this way? How is it that they find it acceptable to hurt the human being who has interacted with them twice a week for 14-16 weeks? And why does it matter that I am a Black woman teaching mostly white students?

Prior to coming to the United States, I had never been referred to as a person of color, much less Black. Race had never been a descriptor of my identity. And it certainly was never a metric of my efficiency or the lack thereof. I was either good at what I did or was not and had to work harder to do better. With the increasing reference to my race and my gender, however, I began to wonder if being female faculty of color equates to presumptions of incompetence. Thus, even though I never thought of myself as a racial category, I soon learned that there was no escaping being multiply marginalized, no matter my qualifications. For example, what do you do with a comment that calls you "very knowledgeable" and caring "about the quality of her students" but "she does not know as much as she thinks she does" or "Women have no place teaching this class" and "she needs to assign less work." It is no wonder that these notions of difference make their way into SETs.

Scholars have found that students' first impression of their professors does influence their perception of that professor and expertise (Littleford et al., 2010). What does that mean for a person like me, who cannot disambiguate my identity? I am a person who, as James Baldwin (1997) once said, puts "my business in the street" (p. 5) the moment I open my mouth! Ironically, Baldwin (1997) observed this of the spoken English of an American living in England. Language, as Baldwin wrote, is revealing of one's "private identity" and "is capable of connecting one with or divorcing one from the larger public" (p. 5). And as a writing instructor, I may not always square my accent with my ability to teach writing because, as Littleford et al. (2010) note, students associate non-white professors with content expertise in racial courses. So, when it comes to students' conceptions of the teaching and the learning of writing and the difficulty faculty may have in presenting written feedback in ways that students perceive as constructive, confusion may ensue. And it may make its way into the hostility, anger, and resistance of SETs, some of which have had me ponder my future in the academy.

Upon going on the job market, for one of the positions I sought, I was asked for, among other things, "evidence of teaching effectiveness as measured by SETs." My heart sank. While I had acquired content and instructional expertise along with a mix of positive and negative SETs, my fixation on the negative ones won the day. No search committee, particularly one that measures teaching effectiveness by SETs, would look at my evaluations and recommend me for hire. Thus, despite my meeting all the required qualifications outlined in the job description, this additional request (not included in the job description but sent out as an

additional screening mechanism) threw me off. Rather than subject myself to the rejection that I was sure would come, I withdrew my candidacy.

Teaching writing is an especially subjective endeavor, particularly when it comes to the iterative process of drafting, feedback, and revision, all of which might privilege the instructor's perspective over that of the student's. See, instructor feedback conveys the idea that there is an unarticulated standard (beyond the rubric) that has not been met. Because, as writers, we identify strongly with our writing, we can see how students' interpretation of the role of the writing instructor may have mismatched expectations about what counts as good writing. And if students' initial impression of the writing instructor presumes incompetence, they may feel the need to punish the instructor's response to their writing with stinging rebukes. Ultimately, such remarks are not a measure of teaching effectiveness because, by all accounts, the instructor is engaging in established pedagogy—the only difference being that they are gendered, racialized, or accented.

As I write this, I have been teaching for over 15 years. I have come to understand SETs as the problem we (academics) all live with (to paraphrase Norman Rockwell). I still get comments such as "She seems to know a lot." As a rhetorician, I cannot help but examine this comment for what it implies. Do students find it objectionable that someone like me, who holds a PhD in rhetoric, knows my subject matter? While my instinct is not to assign any label to these tendencies, I have nevertheless found myself puzzling over whether this comment contains some kind of coded message or if students implicitly doubt my credentials. I am not alone. Research by Smith and Hawkins (2011) shows that identity-based bias in SETs is gender-based bias (Mengel et al., 2017; Rivera & Tilcsik, 2019). Another frequent comment offered as a negative is that I am always prepared and that I come to class on time. I often wish I could go back to the class after I have read these SETs and clear up some things with students. I'd want them to know that I would never have been hired without proof of qualification, that I would not have risen to my rank without meeting the rigorous appointment and promotion processes I have undergone every two years until I earned tenure.

The truth is that academia causes some of us to internalize failure even when we are successful. Through a myriad of ways, the stress of SETs ingrain inadequacy in our lives as academics. The toll they place on our physical and mental health is real, as is their ability to induce stress and self-doubt. On #AcademicTwitter, academics have disclosed feelings of anxiety and general angst at the prospect of reading SETs. Others have decided never to read them at all. For others, SETs exact a tremendous amount of labor. Like others, to deal with harsh SETs, I have diverted time from research to revamp courses, bent over backward to accommodate students' impetuousness, and routinely apologized for students' inability to do their work on time. This constant need to accommodate while proving that I am qualified is enervating.

Seasoned and new college professors who experience the stinging rebuke of SETs have often pondered their (SETs) role in professional development, given that in other professional environments, reviews prepare the reviewee to celebrate some wins and work out approaches to tackle weaknesses. Not so with SETS. Who accounts for the cascading effects of negative ratings and the implications for career trajectory? Having tenure does not inoculate one from the negative effects of SETs, perhaps because of the very human need for validation. And for me, who occupies a place that defines the boundaries of race, gender, and foreignness, the effects can be precarious.

SILVA, FIRST GENERATION IMPOSTOR

If there was a point system for the imposter, I would lose one point for the handful (sometimes two) of low SETs I receive each semester, one point for being a woman in higher education, one point for being part of Generation 1.5, one point for having elementary school teachers who scared my parents into speaking English with us, another point for the spankings I got for leisure reading, and the last point for attending a poor rural high school with science textbooks copyrighted two generations prior. I always had some good excuse for each perfect report card or accolade. Even my becoming a writing professor was sort of a fluke. Students who fail the verbal portion of the SATs don't go on to become writing professors, so I assumed. After two terminal degrees, a master's, two certificates, and, of course, tenure, there is little to apologize for; however, my inability to see myself reflected in my professors and colleagues left me believing that my endless questions and lifelong pursuit for knowledge were evidence of my outsider status.

My parents immigrated to this country in 1965 without a high school education. They labored under the California sun, enduring back-breaking work while raising six children, including my severely handicapped sister. We eventually thrived within the agrarian communities of the valley where immigrant families have long lived in the margins with limited access to literacy resources, such as literacy sponsors, public libraries, and well-staffed and fully stocked schools (Brandt & Deborah, 2001). My academic trajectory as an overachiever paralleled my path as a first-generation daughter raised by austere Catholic parents who did not hesitate to use shame, criticism, and corporal punishment to mold me into a loyal, obedient child. From these parallel worlds emerged my very own internal critic.

Imposter syndrome is chronic across all demographics. According to an article published by the American Psychological Association, imposter syndrome was first coined in 1978 by psychologists Suzanne Imes and Pauline Rose Clance (Weir, 2013; see also Teagan Decker, this volume). They describe it as feeling "like a fraud" and attributing one's success to luck. The phenomenon is most

common among successful high achievers. According to Hutchins and Rainbolt (2016), tenured and tenure-track faculty also experience the imposter phenomenon, in which the highest reported cases were tenure-track faculty. Emotional exhaustion as an outcome of the impostor syndrome was reported by both tenured and tenure-track faculty. It was first believed that only women experienced this problem. In the last two decades, researchers have discovered that men and some minority groups are also impacted significantly by the imposter syndrome. For faculty of color, it is challenging to develop and sustain "a scholarly identity" because faculty often internalize the prejudicial messages about their minority group (Dancy & Gaetane, 2014, p. 367).

Callie Edwards (2019) is a Black female scholar who shares her experiences overcoming the imposter syndrome at a predominantly white male university. She writes:

> Rooted in the ideologies of privilege and oppression, both phenomena ignite a sense of otherness and probate the dominant metanarrative. Whether they feel as though they do not belong (i.e., imposter syndrome) or they feel as though they must prove they belong (i.e., stereotype threat), some marginalized groups are hyperaware of how they are bothered, and this awareness influences how they navigate spaces. (p. 20)

Individuals from marginalized groups often feel they have to hide or alter their true selves to fit in. Since I have worked at predominantly white, affluent institutions of higher education for the past 17 years (i.e., doctoral program and current college), I have had to perform *better* the role of the nurturing, compassionate, accommodating *mother* rather than the strong-willed, firm but understanding matriarch. In my community, women speak their minds in a manner that often appears abrasive to outsiders. The more we care, the more amplified our commitment. The more I care about my students' growth as writers, learners, and thinkers, the less nurturing I may appear. Hence, the more I have to hide my authentic self.

Both Edwards (2019) and I lacked positive representations of female leaders within education, and, similar to Edwards, who had a stuttering problem as a child, I stuttered at times in stressful academic situations. Also, because English was not my first language, I struggled to pronounce certain English words during my primary and secondary education, which made me the butt of many cruel jokes. Even during my college years, I had a male mentor who mocked me when I stuttered. I endured his emotional abuse for a couple of years because I was enamored by his intelligence, confidence, and modest beginnings. Thus, when the occasional SET comment about my pronunciation surfaces, it is a painful reminder of the emotional abuse I endured most of my life regarding the way I speak.

Moreover, writing never came easy for me. It still doesn't. I assume that academics of privilege have fewer doubts than me and labor far less to write. But I followed in my parents' footsteps and did what made the most sense: I worked and worked. For my parents, hard work meant more cows and more land. For me, hard work meant straight A's, a tenured job, and the compulsion to quadruple my disciplinary knowledge to leave no doubts in colleagues, administrators, and students. Failure as a child was not an option if I wanted to avoid abuse, and failure on the tenure track and as a tenured professor was loss of income, access, and opportunity.

For years, since I was a graduate student in predominantly white institutions of higher education, I worked alongside white male colleagues who boasted about neglecting their responsibilities in the classroom while receiving positive SETs. Despite the additional hours I invested weekly to meet with students one-on-one, provide ample feedback on drafts, or develop course materials to scaffold gaps in student knowledge, each semester, there was always a handful of students critiquing my pronunciation, mocking 1-2 grammatical errors on an assignment sheet, or questioning my intelligence. When I tutored and taught in California over 20 years ago, which has a large Latinx community, I mainly received high SET scores, established a positive rapport with students, and felt a strong calling in my work. In recent years, the call to purpose barely persists, like a phantom limb.

It's one thing to endure the criticism of student evaluations in the privacy of your office, but when your department chair, administrative assistant, executive committee, and dean have to read these evaluations each semester, then it's a constant *public* reminder that you DO NOT BELONG here. The experience is tantamount to public shaming. In the humanities, our culture of failure presumes that educators and scholars must be accountable for their individual behaviors and actions. We normalize the daily grind of teaching, research, and writing and only take notice of fellow colleagues when someone publicizes a publication, promotion, or award; otherwise, successes remain private, unlike industries that commonly publicize to internal and external audiences sales reports and share values as indicators of individual and collective success. Failures may be reported similarly, and for many industries (e.g., information communication technology), the culture of failure is changing, in which knowledge workers, influencers, and innovators are encouraged to embrace failure in order to progress and remain competitive in a rapidly changing market. In the humanities, however, our personal failures as writers, researchers, instructors, and community members are kept in the vault of secrecy (Brown, 2015). When we choose to learn from our failures, we may seek the support of a trusted colleague, counselor, or confidant or attend professional development events, but for the most part, Brown states, we internalize our failures as our shame. And our vault of secrets and silence is pried wide open annually when administrators

and fellow colleagues evaluate or question our performance and competence off a single data point—SET.

In a biased system of assessment that privileges certain bodies, personality characteristics, and manners of speech, the only option left for those who do not benefit from such a system is to assimilate. Ibram X Kendi (2019) writes that assimilationist discourse is racist and sexist because we believe that a marginalized group should change and that there is something inherently wrong with this group. If I'm going to succeed and learn from my failures within this biased system, the only option left for me is to assimilate "better" and embody "better" the figure of the nurturing, accommodating professor who speaks "better" standard American English. When the great American success story of the leather-patched academic features values such as hard work, grit, curiosity, and a bit of ingenuity, we naively presume that a life committed to such values would result in success. No amount of hard work, intelligence, or grit equates to higher SET scores. The truth involves a somber acceptance of systemic racism and sexism in higher education manifested in multiple forms, including the standardized evaluation of teaching by students. Most SET surveys are not designed to improve the quality of the course and teaching; rather, they provide an anonymous public platform for students to voice their perceptions, assumptions, and biases about the professor or instructor, regardless of whether they are positive or negative.

The long, arduous journey to tenure nearly broke me. When I first arrived at my current place of employment, I was denied merit pay due to mediocre SETs in non-elective courses, even though I had completed exemplary accomplishments in research and service. When I petitioned my rejection to the department chair, I asked, "Based on the minimum qualifications for merit pay ("excellence" in service, research, AND teaching) and the inherent bias of SETs, how is it possible for me to ever be eligible for merit pay? He dodged the question and simply encouraged me to try again. I never bothered. I knew it was impossible for me to meet their bar of "teaching excellence." After years of steady improvement in my SETs, I put forth my tenure file to the All-College Tenure Committee. My departmental tenure committee was not confident in my ability to acquire tenure and suggested that I withhold my tenure process another year. I felt both rejected and humiliated. I ignored their recommendation and submitted my file anyway to the All-College Tenure Committee, which unanimously approved my tenure file. Although my department intended no harm, their doubts left me questioning my place in the department.

After I received tenure, my department chair raised several concerns about a handful of negative SETs from each course, questioning my competence and commitment to teaching without viewing any of my curricular materials or visiting my classroom. In an email, he provided a rationale for rejecting my research

grant application for $3000, arguing that my research pursuits detracted me from teaching well. The same chair later raised concerns to the Executive Committee about my interests in the Writing Center Director position, doubting my ability to work with students, even though I had successfully collaborated with students multiple times in the past. With each rejection, failure after failure, it was clear my department did not want me there.

A friend once told me, "Now that you have tenure, you don't have to worry about those evals anymore. They can't fire you." True. They can't fire you. But they can deny you merit pay, promotion, research opportunities, and social validation. They can ensure that within a punitive, biased system of assessment, no amount of professional development and self-improvement will free you from feeling like a failure.

NULL, RUNNING AROUND LIKE MOTHER HENS WITH OUR HEADS CUT OFF TO AVOID THE DREADED LOW SETS

As a female professor in an almost all-white and all-female department that prepares students for the majority-female and service-oriented profession of K-12 teaching, I find that even when I do all the work of developing extensive course materials, providing communication through multiple channels, being available almost to the point of being on call 24-7 and identifying and responding to students' needs that are often far beyond the scope of the class—sometimes at the cost of meeting other professional goals—I still risk being penalized with lower SETs if students didn't feel like I did enough to meet their expectations, which are often gendered.

For example, last year, I volunteered to teach an additional one-credit course for incoming freshmen, which required investing a great deal of time into creating a new curriculum, creating a shell on our course management system (CMS), continuously revising the curriculum and CMS shell to adapt to ongoing new requests from the program committee, and working with students to accommodate a variety of needs around attendance, homework completion, accommodations for life events, etc. I'd thought I had managed all of this well until I received course evaluations that weren't just negative but downright snarky. I had brought snacks on the day we did SETs, and one student even complained about the snacks (apparently, she did not like Diet Snapple). I was baffled as I read evaluations that I have never gotten in my other classes, such as, "My teacher was too abrasive and rude for me to develop any feelings stronger than contempt and dislike for her." Or, "My professor was extremely scattered, confusing, and mercurial." I racked my brain to try to identify anything I might have said or done to make one or more students feel this way but was left unsure about when I might have given them that impression. I realized I was being evaluated

on personality, likeability, and delivery rather than on whether I had taught the content the program committee had asked me to teach.

The evaluations made me feel like I had failed on multiple levels. I had failed to teach the course well, failed to read my students' feelings and needs as I'd taught the course, and failed to give my students what they needed. When several students ended the term on academic probation, I wondered whether a negative experience in my class had been a factor, and I felt like I had failed to set them up for a successful college experience. Also, since quite a few of my colleagues had received rave reviews about their courses, I couldn't help but internalize the failure as something that was specifically wrong with me or with my teaching.

Although the sane part of my brain told me that I was probably obsessing more than I should about comments that a couple of eighteen-year-olds might have written without giving it too much thought, I also felt like my college's SET and promotion processes almost required me to worry about the possible ways their words could affect my career. I received these evaluations the term before I was required to submit my post-tenure review and the year before I planned to submit my application to be promoted to Full Professor. The SETs were from a new program that administrators at our college were looking at closely. The negative SETs were from the most recent term, which members of the tenure committee might be most likely to read. Despite twelve years of strong evaluations, I felt like I now had a stain on my record. Plus, I would have to spend extensive space within the two pages allowed for my Teaching Reflection discussing these low reviews, which would give me less space to discuss my numerous teaching accomplishments, including teaching 16 different courses, the innovative partnerships I'd established with middle and high schools, or the fact that I had helped found two master's programs. Colleges' use of SETs for evaluation can often mean that a few negative course reviews risk supplanting years of accomplishment.

It was especially galling that documentable evidence of the quality of my teaching carried far less weight than a few eighteen-year-olds' perceptions and recollections of the course. While students' perceptions of the course are certainly an important data point and are information I use when I re-design courses each term, over-reliance on this data to the exclusion of other data creates a biased and unreliable faculty evaluation system. Even when faculty members can demonstrate evidence of extensive preparation, communication, assessment, and student responsiveness through our CMSs, emails, and logs of discussions with students (35 for that one class of 14 students over a period of 8 weeks), administrators and tenure committee members often rely on SETs because they take less time to read and have the deceptive appearance of being objective data.

SETs create a "customer service" culture in which students expect frequent exceptions and accommodations, and these expectations particularly fall on faculty

within female-dominated fields (Annan et al., 2013). Particularly when I was pre-tenure, I felt compelled to "accommodate" additional needs and expectations in the hopes that my additional work would be acknowledged on my SETs at the end of the term. Instead, I found there was not a clear correlation between doing more and getting better SETs. While saying "no" to additional requests is almost a guarantee of getting a low SET, agreeing to make accommodations that may not have even been asked of a male faculty member is not a guarantee that I will get higher ratings. In fact, the correlation is often negative; students who require the most time and accommodation are more likely to give lower evaluations. The student I bent over backward to help may forget to do the SET, may give me average ratings since they simply expected female faculty members to do this additional work, or may still give me a negative evaluation if they felt like all of my accommodations still weren't accommodating enough. It often feels like the expectations created by SETs leave me with no clear pathway to prove pedagogical effectiveness.

Other scholars have found that what I've described is not atypical for female faculty. Owen (2019) writes, "Investing more time in teaching comes at a price, often decreasing the amount of time available to spend on scholarly activities that are crucial for successful tenure, promotion, and salary reviews." She further explains that constantly striving to meet this higher standard "can encourage a counterproductive downward spiral for some female faculty and faculty of color because it requires those from underrepresented groups to make a greater investment in teaching in order to receive ratings similar to those received by those who aren't subject to this bias." El-Alayli et al. (2018) describe these expectations made of female faculty as "academic momism" (p. 137), in which female faculty are expected to do the additional emotional labor of being more available, nurturing, accommodating, and supportive. In my department, I call this common practice of doing a lot of additional work to maintain the levels of clarity, availability, and responsiveness expected of female faculty "running around like mother hens with our heads cut off." This additional academic and emotional labor can often take up hours (or even days) of additional work each week—time that is often required on an "on call" basis, which further fragments our schedules. These extra time and energy costs can reduce our productivity in other areas and even our quality of life as it eats into our weekends and evenings. Nevertheless, the additional teaching labor required of female faculty is a non-promotable task (Babcock et al., 2022) that is rarely recognized within our institutions and doesn't usually help us with promotion (Hiller, 2020). Given these expectations, while male faculty members may encounter predictable, manageable, and straightforward pathways toward tenure and promotion, female faculty members may encounter an asymptotic curve where it's impossible to fully reach the institution's definition of "successful" teaching, service, and research.

In the case of my optional one-credit course that created such low SETs, I reflected on the many ways I could have done a better job, and I wrote a whole list of ideas for re-designing the course that incorporated my students' feedback. Then, I decided that the benefits of teaching an optional class for a new program were not worth the costs. The extensive time investment in developing the course, making constant changes, and responding to extensive student needs— combined with feeling like I'd been penalized for the effort—made the decision easy. I concluded that other faculty might better be able to serve students in that program.

As a white, tenured professor who is also the child of a professor and from an extended family where everyone has college degrees, I want to acknowledge my privilege navigating academia and acknowledge that I didn't experience the same kinds of significant departmental, administrative, or job search limitations that my co-authors did. Working with my colleagues on this chapter made me realize how frequently female faculty make these types of choices, often in areas that have bigger potential effects on their career trajectories. These decisions often end up creating losses for the institutions that over-relied on SET data and for those institutions' current and future students.

CONCLUSION

Many colleges encourage their faculty to shift their pedagogies to growth and asset models of student achievement, and yet they employ SETs, which can implicitly rely on a deficit model—most strongly for faculty members who are women and/or people of color. Colleges encourage faculty to be thoughtful about gender bias and racial biases and to guard against all the subtle ways they can impact classroom cultures and students' learning, and yet they subject their own faculty who are women and/or people of color to evaluation tools, which they KNOW are subject to racial and gender bias. Moreover, if the feedback faculty members are supposed to use isn't reliable, valid, or actionable, how can faculty members use this feedback to improve their own teaching, and what does this process say about how much the institution really values effective teaching?

SETs create the illusion that teaching can be measured objectively. But the literature shows—and our experiences underscore—that objectivity is a fiction. Kowai-Bell et al. (2012) quote an Education professor's thoughts about SETs, who stated that SETs are not an "objective assessment of instructional skills;" rather, SETs measure *how* students "perceive the teacher makes them feel as a learner and an individual" (p. 348). To persevere in an inequitable system of evaluation that privileges white, heteronormative males and disadvantages, professionally and psychologically, female faculty and female faculty of color (as well as male faculty of color), semester after semester, we are compelled to

rework and edit the narrative script for students so that they *perceive* and *feel* that learning took place, independent of the evidence of meeting learning outcomes. Chan et al. (2014) write that SETs should focus on student learning outcomes "as opposed to improving only the students' perceptions of the teacher" (p. 286). McMurtrie (2024) writes that colleges and universities continue to administer SETs because the "current system is also easy—and cheap." In sum, SET oversimplifies the complexity of student learning and the instructor-student dynamic and reduces it to a numerical value or Yelp review.

Failure can be generative for faculty who have supportive networks and time to reflect on and implement what they've learned and who have the experience and expertise to contextualize failure without feeling overcome by it (Jungic et al., 2020; Laksov & McGrath, 2020; Timmermans & Sutherland, 2020). And failure can be "sideways" if there were paths that led to new and healthy ways of knowing, being, and emoting (West-Puckett et al., 2023). But none of the conditions that make failure a springboard for growth are in place in most colleges' SET systems. Although people learn the most in the short term from successes and from positive feedback (Eskreis-Winkler, 2020), SETs do not consistently highlight successes, nor do they make room for learning. Even when SETs report positive feedback, they are subject to the same biases as negative SETs, such as feedback based on personality and appearance. In response to negative, baffling, or even spurious feedback, faculty may lose motivation or seek to avoid the issue (Eskreis-Winkler, 2020). In addition, the current system, as it is used by most colleges, doesn't usually provide the structures or opportunities for the support, reflection, contextualization, or guidance toward growth that can make failure meaningful and help faculty grow (Jungic et al., 2020; Laksov & McGrath, 2020; Timmermans & Sutherland, 2020). Instead of engaging in a generative meaning-making process, faculty usually end up reading and processing their SETs on their own (perhaps with a glass of wine) and may end up feeling more isolated (Laksov & McGrath, 2020). As Timmermans and Sutherland (2020) wrote of "failure" in academia:

> Failure is individualised and privatised (Gill, 2009). We are called upon to develop resilience—a quality enabling us to withstand the impact of failures and to persevere. However, the burden of overcoming failure is a solitary pursuit and responsibility. We are not further connected to and lifted up by the communities and cultures in which we work. We are not reassured that failure is a normal dimension of being human. (p. 44)

Such isolation can make faculty feel concerned that they are the only ones getting low evaluations or that their lower evaluations indicate deficits in their

teaching ability. It can make SETs feel like something we need to hide, gloss over, mitigate, or even avoid rather than as useful feedback that can help us grow into more effective teachers.

Although managing feelings of failure from low SETs can affect both male and female faculty, the fact that women and people of color (particularly female faculty of color) are more likely to get low SETs—even when they put more time and effort into their courses—suggests that those faculty are more likely to feel isolation, depression, anxiety, and mental exhaustion. The additional time and energy burden imposed by an endless process of improving their courses can negatively affect productivity in other areas, catching women and, particularly, women of color in a "counterproductive downward spiral" in which they are compelled to spend more and more time on their teaching in order to earn scores that their white, male colleagues may be able to earn without so much additional effort (Owen, 2019). Or, as with workplace bullying, continuous negative feedback can cause instructors to disengage, and disengaged instructors aren't as likely to invest extra time or energy innovating their practices, engaging with students or colleagues, or improving their institutions (Hollis, 2015).

When used for hiring, tenure, and promotion, SETs undermine the efforts of universities and colleges to diversify their faculty and teaching staff. Job applicants may preemptively remove themselves from a job search due to worries about low SETs (Walwema), remove themselves from consideration for directing their campus' Writing Center (Silva), or stop teaching an optional course (Null). Consequently, colleges and universities lose the diverse faculty they proclaim to value; female students and students of color lose access to representative faculty members; all students lose exposure to talented faculty who may challenge their thinking and cultivate their skills; and last, our research fields lose the scholarly contributions and narratives of female faculty and faculty of color.

This is not a call to implement SETs better or create a more generative or more reflective process with a cohort of colleagues. This is not about creating better questions on more "observable" behaviors because even those are subject to bias (Boring et al., 2016; MacNell et al., 2015). However, there are other, better ways for instructors to collect and use feedback from students that would be more meaningful. Chan et al. (2014) argue that SETs "should be part of an overall strategic plan that provides reliable triangulated evidence from different perspectives for the improvement of teaching and learning" (p. 286). There are more reliable forms of data instructors can use to demonstrate their teaching effectiveness. For example, faculty members could collect formative student feedback at different points throughout the term as an ongoing reflective practice. One study revealed that this type of formative feedback improved SET scores at the end of the term (Winchester & Winchester, 2014). Formative feedback

allows faculty to apply students' feedback when it can benefit their students; moreover, non-anonymous feedback in the form of student learning reflections offers more meaningful and actionable commentary about the course that benefits students and instructors mutually (Youssef, 2012).

We live in the age of data. Why do colleges rely on a 1960s-era evaluation tool (Wachtel, 1998, cited by Uttl et al., 2016) in the age of email and spreadsheets, plus CMSs that compile data analytics of all course activities (e.g., analytics of site usage by students and instructors, feedback to students, announcements, and email correspondence)? While not all college faculty are comfortable with this level of record retention and even (some might say) surveillance, personnel evaluations that rely on more objective criteria tend to be less biased toward women (Jirjahn & Gesine, 2004). Perhaps one option might be to allow faculty a broader range of choices about what they submit in their evaluation portfolios, such as records of course observations from other instructors, administrators, or faculty; records of interactions with students; student artifacts; course materials; CMS analytics; or other forms of data, including formative and summative course evaluations. Such a process could allow faculty members to provide a more complete picture of what they have achieved and could provide options that could mitigate bias. Hobson and Talbott (2001) broadly define the scholarship of teaching as "the ideology, pedagogy and evaluation of teaching" (p. 26). Based on this definition, the examination of teaching should include a variety of teaching evaluation methods. However, in plenty of instances nationwide, the overreliance on SETs by administrations and search committees has derailed and stalled the career paths of faculty and graduate students.

As long as SETs remain in place, we are forced to endure anonymous attacks, aggregate binders of data to justify our contributions to our department and discipline, knowing full well that our colleagues who benefit from SETs do not have to do the same, and exhaust limited emotional resources and time to appease dissatisfied students and colleagues. If we're going to change this culture of failure that presumes that faculty could learn something of value from discriminating comments about their teaching, it must start with policy changes and drafting a new narrative about teaching and the dialogic and dialectical relationship between faculty and students.

REFERENCES

Annan, S. L., Tratnack, S., Rubenstein, C., Metzler-Sawin, E., & Hulton, L. (2013). An integrative review of student evaluations of teaching: Implications for evaluation of nursing faculty. *Journal of Professional Nursing, 29*(5), 10-24. https://doi.org/10.1016/j.profnurs.2013.06.004

Arthur, L. (2009). From performativity to professionalism: Lecturers' responses to student feedback. *Teaching in Higher Education, 14*(4), 441-454. https://doi.org/10.1080/13562510903050228

Babcock, L., Peyser, B., Vesterlund, L., & Weingart, L. (2022, October 5). Female faculty: Beware the non-promotable task - Mentoring, committee work, and other campus service disproportionately burden women. *The Chronicle of Higher Education*. https://www.chronicle.com/article/female-faculty-beware-the-non-promotable-task

Baldwin, J. (1997). If Black English isn't a language, then tell me, what is? *The Black Scholar: Ebonics*, *27*(1), 5–6.

Baslow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology*, *87*(4), 656-665. https://doi.org/10.1037/0022-0663.87.4.656

Beran, T. N., & Rokosh, J. L. (2009). The consequential validity of student ratings: What do instructors really think? *Alberta Journal of Educational Research*, *55*(4), 497-511.

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research, 0,* 1-11. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

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

Boswell, S. S. (2016). Ratemyprofessors is hogwash (but I care): Effects of Ratemyprofessors and university-administered teaching evaluations on professors. *Computers in Human Behavior*, *56*, 155-162. https://doi.org/10.1016/j.chb.2015.11.045

Bradley, C., & Holcomb-McCoy, C. (2004). African American counselor educators: Their experiences, challenges, and recommendations. *Counselor Education and Supervision*, *43*(4), 258-273. https://doi.org/10.1002/j.1556-6978.2004.tb01851.x

Brandt, D., & Deborah, B. (2001). *Literacy in American lives.* Cambridge University Press.

Brown, B. (2015). *Rising strong: The reckoning. The rumble. The revolution.* Random House.

Buser, T., Noemi, P., & Wolter, S.C. (2022). Willingness to compete, gender and career choices along the whole ability distribution. *Experimental Economics 25*, 1299–1326. https://doi.org/10.1007/s10683-022-09765-8

Chan, C., Luk, L., & Zeng, M. (2014). Teachers' perceptions of student evaluations of teaching. *Educational Research and Evaluation*, *20*(4), 275-289. https://doi.org/10.1080/13803611.2014.932698

Clayson, D. E., & Haley, D. A. (2011, Summer). Are students telling us the truth? A critical look at the student evaluation of teaching. *Marketing Education Review*, *21*(2), 101-112.

Crumbley, D. L., & Reichelt, K. J. (2009). Teaching effectiveness, impression management, and dysfunctional behavior. *Quality Assurance in Education*, *17*(4), 377-392.

Dancy, E. T, & Gaetane J. (2014). Faculty of color in higher education: Exploring the intersections of identity, impostorship, and internalized racism. *Mentoring & Tutoring: Partnerships in Learning*, *22*(4), 354-372. https://doi.org/10.1080/13611267.2014.945736

Davison, E., & Price, J. (2009). How do we rate? An evaluation of online student evaluations. *Assessment & Evaluation in Higher Education*, *34*(1), 51-65. https://doi.org/10.1080/02602930801895695

Edwards, C. W. (2019). Overcoming imposter syndrome and stereotype threat: Reconceptualizing the definition of a scholar. *Taboo: The Journal of Culture and Education, 18*(1), 18-34. https://doi.org/10.31390/taboo.18.1.03

El-Alayli, A., Hansen-Brown, A. A., & Ceynar, M. (2018). Dancing backwards in high heels: Female professors experience more work demands and special favor requests, particularly from academically entitled students. *Sex Roles, 79*, 136-150. https://doi.org/10.1007/s11199-017-0872-6

Eskreis-Winkler, L. (2020, May-June). Maybe failure isn't the best teacher. *Harvard Business Review*. https://hbr.org/2020/05/maybe-failure-isnt-the-best-teacher

Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and Bayesian analyses. *Research in Higher Education*, *53*(3), 353-374. https://doi.org/10.1007/s11162-011-9229-0

Hiller, J. (2020, August 11). After working 50 unpaid hours this week, I fear for my job. *Times Higher Education.* https://www.timeshighereducation.com/blog/im-working-50-unpaid-hours-week-and-i-fear-my-job?fbclid=IwAR27112k45FAPS9NOIhx6FBo3wjRxGNhU7BKTeClDo4TemkJkBdREUAobJo

Hobson, S. M., & Talbott, D. M. (2001). Understanding student evaluations: What all faculty should know. *College Teaching*, *49*(1), 26–31. https://doi.org/10.1080/87567550109595842

Hollis, L. P. (2015, April-June). Bully university? The cost of workplace bullying and employee disengagement in American higher education. *Sage Open, 52*(2), 1-11. https://doi.org/10.1177/2158244015589997

Hutchins, H. M., & Rainbolt, H. (2016): What triggers imposter phenomenon among academic faculty? A critical incident study exploring antecedents, coping, and development opportunities. *Human Resource Development International, 20*(3), 194–214. https://doi.org/10.1080/13678868.2016.1248205

Jirjahn, U., & Gesine, S. (2004). Gender, piece rates and wages: Evidence from matched employer-employee data. *Cambridge Journal of Economics, 28*(5), 683-704. https://doi.org/10.1093/cje/beh027

Jungic, V. J., Creelman, D., Bigelow, A., Côté, E., Harris, S., Joordens, S., Ostafichuk, P., Riddell, J., Toulouse, P., & Yoon, J. (2020). Experiencing failure in the classroom and across the university. *International Journal for Academic Development*, 25(1), 31-42.

Kendi, I. X. (2019). *How to be an antiracist*. One World.

Kornell, N., & Hausman, H. (2016, April 25). Do the best teachers get the best ratings? *Frontiers in Psychology*, 7, 570. https://doi.org/10.3389/fpsyg.2016.00570

Kowai-Bell, N., Guadagno, R. E., Little, T. E., & Ballew, J. L. (2012). Professors are people too: The impact of informal evaluations of professors on students and professors. *Social Psychology of Education*, *15*(3), 337-351. https://doi.org/10.1007/s11218-012-9181-7

Laksov, K. B., & McGrath, C. (2020). Failure as a catalyst for learning: Towards deliberate reflection in academic development work. *International Journal for Academic Development, 25*(1), 1-4. https://doi.org/10.1080/1360144X.2020.1717783

Lindahl, M., & Unger, M. (2010). Cruelty in student teaching evaluations. *College Teaching*, 58(3), 71–76.

Littleford, L. N., Ong, K. S., Tseng, A., Milliken, J. C., & Humy, S. L. (2010). Perceptions of European American and African American instructors teaching race-focused courses. *Journal of Diversity in Higher Education*, *3*(4), 230-244. https://doi.org/10.1037/a0020950

MacNell, L., Driscoll, A., & Hunt, A. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Journal of Collective Bargaining in the Academy*, Article 52. http://thekeep.eiu.edu/jcba/vol0/iss10/52

McMurtrie, B. (2024, February 6). Teaching evaluations are broken. Can they be fixed? *The Chronicle of Higher Education.* https://www.chronicle.com/article/teaching-evaluations-are-broken-can-they-be-fixed

Mengel, F., Sauermann, J., & Zölitz, U. (2017, September). Gender bias in teaching evaluations [Discussion Paper Series, No. 11000]. Institute of Labor Economics (IZA), Bonn.

Miles, P., & House, D. (2015). The tail wagging the dog: An overdue examination of student teaching evaluations. *International Journal of Higher Education 4*(2), 116-126. https://doi.org/10.5430/ijhe.v4n2p116

Mitchell, K. M., & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics*, *51*(3), 648-652. https://doi.org/10.1017/S104909651800001X

Owen, A. (2019, June 24). The next lawsuits to hit higher education. *Inside Higher Ed. https://www.insidehighered.com/views/2019/06/24/relying-often-biased-student-evaluations-assess-faculty-could-lead-lawsuits-opinion?utm_source=naicu*

Pittman, C. T. (2010). Race and gender oppression in the classroom: The experiences of women faculty of color with white male students. *Teaching Sociology*, *38*(3), 183-196. https://doi.org/10.1177/0092055X10370120

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education*, *3*(3), 137. https://doi.org/10.1037/a0019865

Rivera, L. A, & Tilcsik, A. (2019). Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. *American Sociological Review*, *84*(2), 248-274. https://doi.org/10.1177/0003122419833601

Robinson, J. (2018). *Through their eyes: A grounded theory study of resilience for Black women in counselor education* [Unpublished doctoral dissertation]. The University of Texas at San Antonio.

Roseboro, D. L. (2021). Dismantling the architecture of "good" teaching. In L. U. Taylor (Ed.), *Implications of race and racism in student evaluations of teaching* [Kindle edition] (pp. 43-66). Lexington Books.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures?

Assessment & Evaluation in Higher Education, 25(4), 397-405. https://doi.org/10.1080/713611436

Smith, B. P., & Hawkins, B. (2011). Examining student evaluations of Black college faculty: Does race matter? *Journal of Negro Education 80*(2), 149-162.

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

Spooren, P., & Christiaens, W. (2017). I liked your course because I believe in (the power of) student evaluations of teaching (SET). Students' perceptions of a teaching evaluation process and their relationships with SET scores. *Studies in Educational Evaluation*, 54, 43-49. https://doi.org/10.1016/j.stueduc.2016.12.003

Sprague, J., & Massoni, K. (2005). Student evaluations and gendered expectations: What we can't count can hurt us. *Sex Roles*, *53*(11–12), 779–793.

Storage, D., Horne, Z., Cimpian, A., & Leslie, S. J. (2016). The frequency of "brilliant" and "genius" in teaching evaluations predicts the representation of women and African Americans across fields. *PloS One*, *11*(3), Article 0150194.

Taylor, L. U. (2021). Their voices must be heard. In L. U. Taylor (Ed.), *Implications of race and racism in student evaluations of teaching* [Kindle edition] (pp. 17-40). Lexington Books.

Timmermans, J. A., & Sutherland, K. A. (2020). Wise academic development: Learning from the 'failure' experiences of retired academic developers. *International Journal for Academic Development*, *25*(1), 31-42.

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.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, *23*(2), 191-212, https://doi.org/10.1080/0260293980230207

Weir, K. (2013). *Feel like a fraud*. American Psychological Association. https://www.apa.org/gradpsych/2013/11/fraud

West-Puckett, S., Caswell, N. I., & Banks, W. P. (2023). *Failing sideways: Queer possibilities for writing assessment.* Utah State University Press.

Winchester, T. M., & Winchester, M. K. (2014). A longitudinal investigation of the impact of faculty reflective practices on students' evaluations of teaching. *British Journal of Educational Technology, 45*(1), 112-124. https://doi.org/10.1111/bjet.12019

Yao, Y., & Grady, M. L. (2005). How do faculty make formative use of student evaluation feedback? A multiple case study. *Journal of Personnel Evaluation in Education*, *18*(2), 107-126. https://doi.org/10.1007/s11092-006-9000-9

Youssef, L. S. (2012). Using student reflections in the formative evaluation of instruction: A course-integrated approach. *Reflective Practice*, *13*(2), 237-254. https://doi.org/10.1080/14623943.2011.626031