An overview of research on use of student surveys to evaluate teaching

Doug Ward

Associate Director, Center for Teaching Excellence

Associate Professor, School of Journalism and Mass Communications

University of Kansas

Last updated November 2021

An overview of research on student surveys

Introduction and overview

Researchers into student surveys of teaching have come to widely varied conclusions about the validity of the surveys, the insights they provide, and the biases they do or do not contain. One of the few points researchers agree on is that **student surveys should be only one measure of teaching effectiveness**.

Student ratings provide a limited and narrow view of a course and fail to capture the types of evidence-based strategies that improve student learning (National Academies of Sciences, Engineering, and Medicine, 2020). They generally do a good job of measuring student satisfaction but "do not measure *directly* how much or how well a class of students has learned or any other aspect of achievement" (Abrami, d'Apollonia, and Rosenfield, 2007, p. 394). They also do not "match or measure the full range of academic functions nor ever-increasing obligations of faculty" (Wallace, Lewis, and Allen, 2019, p. 9). Nor do they reflect the innovative approaches that instructors have taken to improve student learning, especially during extraordinary circumstances like the Covid-19 pandemic.

KU's own policies reflect the need for multiple sources of evidence about an instructor's teaching. Even so, the use of multiple measures is widely ignored, leading to a loss of trust in the evaluation system (Austin, Sorcinelli and McDaniels, 2007). That became readily apparent early in the pandemic as faculty demanded that student survey results be excluded from personnel files, arguing that students would not account for the unusual circumstances brought on by the pandemic or understand the significant work involved in shifting to remote teaching. Those concerns underscore the need to follow university policy and broaden the approach used to evaluate teaching.

The research literature also makes clear that **student surveys of teaching should be grounded in a shared understanding of good teaching**. They rarely are. Students, faculty and administrators often have differing views, and criteria vary widely across and within schools, departments and disciplines (Abrami, d'Apollonia, and Rosenfield, 2007). KU has no single definition at the university level, although questions on student surveys send a message about what should be valued.

This overview of literature highlights key areas of agreement, disagreement and concern about student surveys of teaching. It draws from a wide range of literature and is intended to provide an overview of the scholarly thinking about student surveys. It is far from comprehensive, and it is *not* intended to argue against the use of student surveys of teaching. The student voice is a crucial component in the evaluation of teaching. An earlier version of this literature review provided context and guidance for the Task Force on Student Surveys of Teaching in 2020-21, and is now intended to provide context to the changes that were made in KU's standard survey starting in Spring 2021.

A note about definitions

This paper uses the term *student surveys of teaching* to refer to the end-of-semester surveys that gather student feedback about instructors and college courses. These surveys are sometimes known as *student evaluations of teaching*, *student ratings*, and *course evaluations*, among other terms. *Student surveys of teaching*, or just *student surveys*, seems the most accurate and neutral term.

Background and context

Student surveys of teaching originated in the early 20th century, and their use spread widely in the 1960s and 1970s as students demanded a voice at their institutions. The surveys were used primarily to help faculty improve their teaching until the 1970s, when administrators began applying them to personnel decisions (Galbraith, Merrill, and Kline, 2012). Clayson (2021, p. 3) says the purpose of the surveys in the 1960s and 1970s varied widely, with "no consensus about what the process was designed to measure." Spooren et al. (2013) say that the use of student surveys for faculty evaluation arose from an increasing emphasis on "quality assurance," "performance management" and "consumer satisfaction" in universities. Similarly, Cox, Rickard, and Lowery (2021) attribute the elevation in the importance of student surveys to higher education's adoption of a consumerist model. That model, they say, leads students to pressure instructors for higher grades and to threaten to leave poor comments on surveys if instructors fail to comply. Titus (2008, pp. 413-414) goes even further, saying that student surveys grew in use and power as part of an "accountability agenda" that followed a "devaluing of faculty."

Research into student surveys of teaching has been conducted nearly as long as the surveys have existed, and Emmelman and DeCesare (2007, p. 228) describe the literature as "truly vast." As the stakes in using the surveys increased, so did the volume of research seeking to determine their validity, biases, and appropriate use (Galbraith, Merrill, and Kline, 2012). Those aspects have never been fully determined, because of widely varying approaches to student surveys and because of a lack of widely accepted criteria for high-quality teaching (Clayson, 2009). Emmelman and DeCesare (2007, p. 228) say a ratings scale has been used solely for standardization purposes. Hamilton (1980) says concerns about student satisfaction and educational quality perpetuate the use of student surveys, with senior faculty tending to trust the ratings more than junior faculty. Michael Scriven, a pioneer in the scholarly pursuit of evaluation, describes evaluation as "the determination of the value, merit, worth or significance of things" (2015, n.p.). Evaluation, he says, should measure "performance against goals" but also "procedures for the evaluation of goals" (Scriven 1966, p. 18). No matter how it is conducted, though, "Evaluation scares people," Scriven (2015, n.p.) says.

Researchers in education have been among the biggest proponents of student surveys to evaluate teaching and have conducted much of the research into their use. Clayson (2021, p. 5) says much of the early research came from a relatively small group of researchers who had "a philosophical and practitioner affinity toward the idea of student evaluation of instruction." The validity of much of the research before about 2000 has been questioned on methodological and statistical grounds, though, and researchers in other fields have generally found research into the surveys' validity lacking (Clayson, 2009). Clayson (2021) says that student surveys of teaching have also failed to adapt as student attitudes changed, writing: "What students appreciate in an instructor in one generation might not be the same for students of another era" (p. 6). Even those who support the validity of student surveys of teaching (Abrami, d'Apollonia, and Rosenfield, 2007) say that evidence of effective teaching should come from multiple sources and that "student ratings of specific teaching dimensions should not be used indiscriminately for summative decisions about teaching effectiveness" (p. 389).

Increasing scrutiny

Over the past few years, the use of student ratings as a means of evaluating teaching has come under increasing scrutiny, with some universities doing away with student surveys altogether, some renaming them to avoid use of the term "evaluation," and others increasing efforts to expand the use of evidence beyond student surveys (National Academies of Sciences, Engineering, and Medicine, 2020). A consortium of organizations that includes the Association of American Universities, the National Academies of Sciences, Engineering and Medicine, Accelerating Systemic Change in STEM Higher Education, the Bay View Alliance, the Network of STEM Education Centers, and an NSF-funded organization known as TEval (which involves the Center for Teaching Excellence at KU) has worked to raise awareness of the need for a fairer, more nuanced approach to evaluating teaching. (The consortium held a conference on the evaluation of teaching in Fall 2019 and another in January 2021. TEval sponsored another series of national discussions in October 2021.) The AAU, the Cottrell Scholar program, and the Research Corporation for Science Advancement say that established practices of evaluating teaching (primarily use of student survey data reduced to a numerical scale) have impeded efforts to improve teaching at research universities and have led to an undervaluing of teaching in general (Dennin et al., 2017).

A large percentage of faculty members say that teaching should be evaluated with the same seriousness as research (Flaherty 2015), and Hutchings, Huber and Ciccone (2011) argue that a more nuanced approach to evaluating teaching could lead to increased emphasis on teaching in the university rewards systems. Few conversations about student surveys of teaching involve doing away with them, though; rather, the emphasis has been de-emphasizing them as a sole or major source of evaluation and providing a fairer system that reflects instructors' efforts at improving student learning and adopting evidence-based teaching practices. Richardson (2005) says instructor use of student survey data varies widely, in part because there is little or no incentive to do so, and little guidance on how to make sense of the survey results. To have the most validity, Richardson says, survey results should be published so that students can see that institutions value their opinions. Clayson (2021, p. 144) calls the current use of student surveys "deeply flawed" and says that one could assume that use of student surveys of teaching "has become more of an ideological and bureaucratic necessity than an evidence-driven one."

Validity of student surveys

Many researchers say that the validity of student surveys has been proved and that further research should focus on improving questions and responses. One of the most widely cited researchers (Marsh, 2007) calls student ratings multidimensional and reliable but warns that a global rating "cannot adequately represent the multidimensionality of teaching." Marsh says validity depends on "a continual interplay between theory, research and practice" (p. 327). Benton and Cashin (2011) argue that student surveys tend to be statistically reliable, valid, and relatively free from bias. Wilson (1988, p. 79) says that not only are student surveys valid but that they "can be considered an advance, an innovation, a means to improve the academy." And yet, he says, the surveys reinforce the status quo and hinder meaningful change. Marsh (2007) says, though, that there are still disputes about whether student surveys "measure effective teaching or merely behaviors or teaching styles that are typically correlated with effective teaching" (p. 322).

In a meta-analysis of research from the 1950s to the mid-2000s, Clayson (2009) found little or no link between learning and student surveys of teaching, and he suggested that rigor in a class led to lower ratings on surveys. That is, students who perform well on exams that require memorization often perceive greater learning and thus give higher scores to instructors. Those in classes requiring conceptual, abstract or analytical thinking often give lower scores. Clayson warned against using scores from student surveys as a method of comparing instructor performance. Similarly, Onwuegbuzie, Daniel, and Collins (2009) argue that their examination of multiple components of validity in student surveys of teaching raises "serious doubt" about the use of the surveys for evaluating teaching.

In an analysis of 12 years of scholarship, Spooren, Brockx, and Mortelmans (2013, p. 599) say that researchers have failed to provide clear answers about the validity of student surveys to evaluate teaching and that the tension between their use as formative and evaluative tools makes them "fragile." They say a new approach may be needed because of a shift toward student-centered teaching and a repeated use of the surveys over a students' time in college (which may sour students' views on the surveys). Artz and Welsch (2013) found that higher course evaluations were generally associated with higher student GPAs but that students' grades generally declined in subsequent years after taking courses with highly rated professors. Clayson (2021, p. 108) says validity of student surveys is difficult to determine because there is "no widely accepted definition of what process is intended to be measured." In another survey, Clayson and Haley (2011) found that substantial portions of students purposely provided false information in rating their instructors or purposely made false comments about instructors.

In a more recent meta-analysis, Uttl, White and Gonzalez (2016) argue that previous meta-analyses arguing in favor of student surveys failed to account for small sample sizes and contained methodological weaknesses that led to faulty findings. Those earlier studies argued that student surveys were strongly connected to teaching effectiveness and student learning. Instead, Uttl, White and Gonzalez say that the use of student surveys is based on unrealistic notions that a few questions that students answer at the end of a semester can measure teaching effectiveness. Despite decades of research, they say, there is no evidence that students learn more from professors who receive higher ratings. Spooren, Brockx, and Mortelmans (2013, p. 599) say the surveys run the risk of "equating student *opinions* with *knowledge*" and they say that student surveys are often disconnected from students' perspectives of effective teaching, thus diminishing their validity. Stroebe (2016) says the high-stakes use of student surveys in personnel decisions has led to leniency in grading, with the surveys essentially becoming measures of consumer satisfaction.

Uttl, White and Gonzalez (2016) say that student surveys provide little more than a view of student satisfaction. Wilson (1988) goes even further, arguing that student surveys hinder change by promoting student passivity and portraying education as a mere transfer of information. The surveys are devoid of critical thought about teaching and learning and reinforce "a concept of educational improvement as an idiosyncratic behavior: If only the individual teacher would reform, correct his or her inadequacies, and bring them in line with the ideal teacher posited by the form, then significant improvement would result" (p. 88). Similarly, Titus (2008, p. 414) calls student surveys tools of conformity that generate "an illusion of objectivity," perpetuate passive learning, and punish instructors whose teaching "deviates from what is officially textually recognized as good practice." Clayson (2021, p. 71) says the lack of association between survey scores and learning "presents a serious challenge to the validity of the evaluations."

Biases in student surveys

From the 1980s to about 2010, researchers found little difference in student ratings scores between male and female instructors (Li and Benton, 2017). Laube et al. (2007) challenge those findings, saying that the results are far more mixed, especially because most of those studies did not account for gender roles or grades. Quantitative measures often mask an underlying gender bias and are based on the assumption that a particular numerical rating is consistent in all contexts, they say. Cox, Rickard, and Lowery (2021) call student surveys a measure "of customer satisfaction, not a measure of learning, teaching effectiveness, or teaching quality" (p. 10). They say that universities should eliminate Likert scale ratings in favor of yes/no answers, which, they say, could cut down on biases and encourage students to be more truthful in their responses.

Valsan and Sproule (2008) contend that student survey scores are tied to students' cultural beliefs and expectations and thus have no validity in measuring the quality of teaching. Students are not impartial observers, they say, and the anonymity of the surveys punishes instructors who fail to meet student expectations and leads to collusion between the instructor and students. They call this a "lethal flaw" (p. 946). Wallace, Lewis, and Allen (2019) conclude that women and faculty of color receive more negative or derogatory comments than men do. Relatedly, they say that administrators view ratings in inconsistent ways. Low scores for a white, male instructor are often seen positively if students complain of overly high expectations or too much reading. The same scores and complaints are often seen negatively if the instructors are women or faculty of color. They warn against overuse of students' written comments, saying that those comments are often reactionary rather than constructive. Hu (2021) goes even further, saying that “it is impossible for any kind of education and teaching evaluation method to be absolute, objective and accurate” (p. 311).

Feldman (2007) says there is no consensus on a definition of bias in student ratings and no evidence that the gender of students affects student survey scores. He also says the literature has found few statistical differences in the scores of male and female professors. When there was a difference, it favored the score of female instructors. Feldman calls the belief in gender bias in student surveys of teaching a myth. Arnold and Versluis, (2019) similarly found no evidence of gender bias in student surveys. Li and Benton (2017) say that questions that use neutral language can generally mitigate gender bias, especially if multiple sources of evidence are used. Other researchers argue that that student surveys are easily manipulated and are often tied to an instructor’s reputation.

Gender bias

Researchers have raised concerns about potential biases in student surveys of teaching since at least the 1980s (i.e., Martin, 1984). Those concerns have gained increasing traction in the past several years as a growing number of studies has raised doubts about student surveys’ bias against women (e.g., Peterson et al., 2019; Mengel, Sauermann, and Zölitz, 2019; Fan et al., 2019; Boring, 2017). In a review of literature, Binderkrantz, Bisgaard, and Lassesen (2021) point to similar biases in the evaluation of public sector employees, principals’ evaluations of teachers, and in patient satisfaction with doctors.

Martin (1984) argues that teaching is seen as a traditional female activity that is valued less than research, which is seen as a traditional male activity. Women often spend more time on teaching and committee work than men do (Martin, 1984). Baldwin and Blattner (2003, p. 28) say, though, that college teaching is still perceived as a “male occupation.” Baldwin and Blattner (2003) describe student ratings as overly powerful and under-examined. We can’t assume students will provide fair evaluations, they say, because “too much evidence and too many horror stories exist among faculty to safely assume that students are immune from influences that detrimentally affect the ratings that they give” (p. 30). Lord (2008) says that survey results have led to promising faculty leaving academia, to instructors inflating grades to keep their jobs, and to universities abandoning learning for student satisfaction.

Researchers are far from universal in their thoughts about gender biases in student surveys of teaching. In a substantial review of literature, Benton and Cashin (2011) say there is little or no evidence to suggest widespread bias against women in student surveys of teaching. Similarly, in two experiments, Binderkrantz, Bisgaard, and Lassesen (2021) found no signs of bias in student surveys and raise the possibility that women receive higher scores than men. In an analysis of meta-analyses, Wright and Jenkins-Guarnieri (2012) discount the likelihood of gender bias and conclude that student surveys are valid as a feedback tool as long as they are combined with additional input like consultations with peers or administrators.

Mitchell and Martin (2018) argue, though, that students use different criteria to evaluate women than they do men, focusing on such things as personality, appearance, and perceptions of intelligence. Because the professoriate is male-dominated, women are assumed to have lower ranks, with men seen as professors and women seen as teachers. Similarly, Martin (1984) says her interpretations of studies of student surveys suggests that female faculty with masculine teaching styles (authoritative) receive higher ratings but that those with feminine styles (warmth and supportiveness) are more effective in helping students learn. Baker and Copp (1997) say that students often hold contradictory and unrealistic expectations for female faculty members. When female instructors fail to meet students' gendered expectations, scores on student surveys decline. This is amplified in larger classes. Colleagues on promotion and tenure committees likewise take gendered views of instructors and take student ratings at face value, with instructors becoming "one-dimensional characters" who need no deeper interpretation (p. 42).

In an analysis of student surveys from business and economics courses in the Netherlands, Mengel, Sauermann, and Zölitz (2019) found that junior instructors who are women received lower scores than their male colleagues, primarily because of lower scores given by male students. The scores were even lower for women who taught large math-related courses. They found little difference in scores among female and male senior faculty, but they say gender bias is "sizeable and systematic" (p. 537). In an analysis of student surveys at a French university, Boring (2017) found that female professors received lower scores on student surveys than their male colleagues did, in large part because male students gave significantly higher scores to male professors. As a result, Boring argues, female instructors may spend considerably more time on time-consuming teaching activities in an effort to increase their survey scores, hindering their chances of advancement. Boring concluded that students may simply be unable to assess the teaching effectiveness of their instructors. Fan et al. (2019), in an analysis of more than 500,000 student surveys in Australia, argue that survey scores are a poor measure of teaching performance because of widespread bias against women and instructors whose native language is not English. The effects of bias were greatest in science and business courses, although that bias was reduced when a department's faculty was more diverse. They found no significant gender bias in arts and social science courses, but significant bias against faculty whose first language is not English. The overall biases were the same for undergraduate and graduate courses.

Peterson et al. (2019) argue that gender bias in student surveys of teaching is primarily implicit, something that happens automatically and without awareness. Offering a prompt that makes students aware of this potential bias can have a positive effect on survey results, their research suggests. Similarly, Clayson (2021, p. 15) says that "gender bias appears to be culturally related and is more subtle than assuming one gender will uniformly receive higher evaluations."

Racial and ethnic bias

Laub et al. (2007) say that students generally assume that white, male professors have authority but don't automatically view female professors or faculty of color in that same position of authority. Smith and Hawkins (2011) found that Black faculty members in a predominately white research university received lower scores than their white counterparts, especially on questions related to teaching ability and value of a course. Pittman (2010) says that white, male students are far more likely than other students to challenge the authority of female faculty of color, question their competence and disrespect their scholarly expertise. This often leads to lower course ratings. Randolph (2010) found much the same, saying that a single disgruntled white student can spread mistrust of and anger toward a Black instructor, turning student surveys into "mob action resulting in power by force" (p. 132).

Parker and Neville (2019) say that white students often come to college from predominately white high schools and have little previous interaction with Black or Latino authority figures, something that can lead to criticisms of teaching styles, questioning of academic expertise, and lower scores in student surveys of teaching. Croom (2017) argues that racist and sexist attitudes lead to lower ratings for women of color, and that universities' emphasis on survey ratings has contributed to a marginalization of Black women in academia, making it difficult for them to move beyond the associate professor level or even to gain a tenure-track job.

Arnold and Versluis, (2019) found that students from cultures where citizens are more likely to defer to those in power (like China) tend to give higher scores on surveys of teaching than students from countries where citizens are more likely to challenge authority (like the U.S.). They also argue that students from collectivist cultures are more likely to give lower scores to instructors who focus on individual achievement.

Response rates and selection bias

A shift to online collection of student surveys of teaching has led to substantial declines in participation rates at colleges and universities across North America. One study (Groen and Yves, 2017) estimates participation rates of about 50% at most universities, although that seems optimistic post-pandemic. For example, response rates at the University of Saskatchewan average 21%.[1] For courses of 200, the Center for Teaching and Learning Enhancement suggests that a 15% to 25% response rate is adequate. For smaller courses (30 students), it recommends a response rate of 40% to 53%.

Several researchers have questioned the validity of low response rates. Goos and Salomons (2017) argue that low and variable response rates skew the results of student survey scores, making any comparison of scores across courses or departments problematic. Trying to adjust for this is difficult, they say, because the bias is based on unobserved characteristics. They found, though, that women were more likely than men to complete student surveys, as were students who earned higher grades. Students who feel stressed or overloaded are less likely to complete online course surveys, which they view as placing additional demands on them at a particularly stressful time (Young et al., 2019). Alhija (2017, p. 1) says that low response rates make student survey results unrepresentative and argues that surveys "fall short from being a significant basis for

[1] Interview with director, June 2022.

informing and improving teaching." Similarly, Nowell, Gale, and Handley (2010) say that the switch to online surveys casts down on the surveys' validity, especially mean scores, given a smaller sample size and an increased variance in responses. Richardson (2005, p. 406) says that students who complete course surveys are "systematically different" from non-respondents in both attitudes and experiences, creating a bias that cannot be accounted for through statistical weighting. This bias can only be minimized by increasing participation in the surveys. They say, though, that requiring participation in class raises ethical issues in that research guidelines stipulate that participants be allowed to withdraw at any time.

Feistauer and Richter (2017) suggest that a course needs at least 25 students for ratings to be considered reliable. They say that survey results are more reliable in instructor-centric courses like lectures than in seminars or courses in which students play a larger role. They also say that factors such as room size, room location, topic of the course and size of the class can play a role in ratings. Zipser and Mincieli (2018) found that a university's switch to online gathering of student surveys, combined with a wording change, led to a decrease in average scores of 0.14 to 0.25 of a point. Extending the time that students had to complete the surveys had no effect on ratings. They warned, though, that low response rates were associated with selection bias. Wolbring and Treischl (2016) argue that biases created by student self-selection in completing the surveys and the ease of manipulating survey results through various means of rewarding students for positive feedback makes them poor tools for comparing different courses or instructors. They strongly urge universities to avoid student survey results as a tool for performance evaluation. Fosnacht et al. (2017) argue, though, that low response rates do not necessarily lead to bias, although their study said that a valid study needed at least 50 respondents. That would suggest that lower response rates would matter less in larger classes but would make those in small classes highly suspect.

Chavez (2021) found that students who were intrinsically motivated were more likely to complete course surveys. Those who were extrinsically motivated were more likely to skip course surveys and use Rate My Professor. Similarly, Young, Joines, Standish and Gallagher (2019) say that students who feel engaged in a class are more likely to respond to online surveys of teaching and that having students complete surveys in class may increase participation rates but also diminish the volume of open-ended comments, which students often forgo because of time restraints.

Use of a numerical scale

Concerns about the use of student surveys to evaluate faculty teaching have been prevalent since at least the early 1970s (Rodin, 1973a). Rodriguez (2019) argues that administrators can interpret scores on student surveys in whatever way they choose, "weaponizing" them to force instructors to adhere to a particular grade distribution, to insist on arbitrary increases in ratings, and ultimately to push out faculty who are "different" and who fail to conform to the thinking of mostly white senior colleagues. Rodriguez, Rodriguez and Freeman (2020) argue that student surveys are part of a "fetishization of numerical data" that reflects a white, male perspective and reinforces biases against underrepresented faculty. They equate student surveys to the use of phrenology in the 19th century, when scientists used measurements of skull sizes to validate what they already "knew" about the superiority of whites. Esarey and Valdes (2020, p. 1106) recommend extreme caution with student surveys, saying that using numerical results "to identify poor teachers can result in an unacceptably high error rate even under the most optimistic scenarios supported by empirical research." That is because what they call irrelevant influences (such as faculty appearance) enter into survey results no matter how carefully a survey is designed. They recommend using multiple sources of evidence but say that survey results could also be used to flag potential problems for further analysis of instructor teaching.

Galbraith, Merrill and Kline (2012, p. 370) warn against use of student ratings in personnel decisions, calling that practice "counterproductive, if not dangerous." Laub et al. (2007) argue that eliminating a global numerical ratings is one way to mitigate bias against faculty of color because a few hostile students can distort numerical ratings. Clayson (2009) warns against using scores from student surveys as a method of comparing instructor performance. Feistauer and Richter (2017) argue say that student characteristics such as personality, competence and tendencies to rate all instructors as either positively or negatively have a considerable effect on student ratings, arguing that "student evaluations cannot be regarded as pure measures of teaching quality" (p. 1276). Rodin (1973a) points to many weaknesses in the way validity in student surveys is measured, arguing that students use widely varied criteria for rating instructors and that administrators generally look only at mean scores.

Gray and Bergmann (2003) call student surveys a blunt instrument incapable of making fine distinctions among instructors. They say that use of a mean score on student surveys has allowed administrators to punish anyone who falls below that mean. That approach punishes good instructors in departments filled with excellent teachers and rewards poor instructors in departments that have mostly bad teachers. It deters innovation and undermines instructors' expertise, leading to a system that is "inaccurate, misleading, and shaming" (p. 46). An arbitrator in a case brought by faculty at Ryerson University in Canada is equally blunt, calling the use of average scores in student survey results "fundamentally and irreparably flawed" (Ryerson University v Ryerson Faculty Association, 2018). The American Sociological Association (2019) says questions on student surveys "should focus on student experiences" and "be framed as an opportunity for student feedback, rather than an opportunity for formal ratings of teaching effectiveness." Ali, et al. (2021, p. 5) say that use of scales and opinion-focused questions can dilute the usefulness of student surveys by invoking "spontaneous responses rather than deep and thoughtful reflections"

Other biases

In a wide-ranging study of the research into student surveys of teaching, Clayson (2021) concludes that a halo effect clearly exists. That is, students tend to give higher scores to younger instructors and to instructors they perceive to be more physically attractive. Additionally, students who receive chocolate, cookies or other treats on the day surveys are administered tend to rate instructors higher. Research also calls into question the truthfulness of some students who give either higher or lower scores to instructors they either like or dislike. Clayson (2021) says students generally answer survey questions based on an overall feeling about a class or an instructor, giving instructors a similar grade to what they expect to receive. Those who think of themselves as customers at their college or university are more likely to provide false information (Cox, Rickard, and Lowery, 2021). Galbraith, Merrill and Kline (2011) argue that student surveys have little or no connection to effective teaching or student learning.

Guidelines and findings in brief

What follows is a compendium of best practices, advice and guidance on the use of student surveys of teaching, along with brief descriptions of the type of bias researchers have found. These areas of bias have not been proved universally; they are simply areas where some researchers have identified bias in student surveys of teaching. In many cases, researchers have found conflicting evidence (see, for instance, Wallace, Lewis, and Allen, 2019); Benton and Cashin, 2011). Even so, evidence about the weaknesses and biases in student surveys of teaching have mounted over the past 20 years, suggesting that they should not be used as the primary or sole factor in evaluating an instructor's teaching.

What students are generally qualified to judge

- What occurred in the class, including organization, use of class time, and the approaches an instructor took to help students learn
- Clarity of goals, expectations and presentation
- Timeliness and clarity of feedback
- Availability of instructor outside class
- Sense of class climate
- Sense of workload compared with other classes
- How often a class engaged in discussion
- Quality of an instructor's presentations

Sources: Clayson (2021); Benton and Young (2018); Task Force on the Assessment of Teaching and Learning (2007); Frey (1974).

What students are generally NOT qualified to judge

- Quality of course content
- Instructor's knowledge of subject matter
- Effectiveness of course design or effectiveness of the instructor
- Appropriateness of course goals
- Quality of the instructor's assessment of students

Sources: Clayson (2021); Benton and Young (2018); Task Force on the Assessment of Teaching and Learning (2007); Frey (1974)

Areas of evaluation that instructors and peers can provide

- **Instructors**. Reflection; evidence of course modifications and self-improvement; examples of syllabi and course materials; evidence of student learning, including examples of student work and factor analysis of rubrics; evidence of use of effective teaching practices, creation of an inclusive climate, use of appropriate materials; explanation of extenuating circumstances. (See Benchmarks for Teaching Effectiveness for additional elements.)

- **Peers**. Knowledge about discipline, pedagogy and specific practices. Generally valid as long as peers take the time to evaluate materials and provide honest feedback. Unreliable if there is no shared vision of good teaching or if peers rely on a single class visit without reviewing class and instructor materials. Review of evidence of student learning, as well as teaching practices, course materials, course design, instructor preparation, professional development, involvement in teaching community. (See Benchmarks for Teaching Effectiveness for additional elements.)

Practices that improve the effectiveness of evaluation of teaching

- **Evidence from multiple sources**. Approaches that rely too much on a single measure leads to loss of trust in a system.
- **Evidence that is compared over time** and identifies trends for an individual instructor. Evidence includes authentic measures that show what a student can do (portfolios or examples of student work, rather than exam scores).
- **Feedback that is timely and focused**. Formative feedback to instructors should be timely, focus on one course at a time, identify strengths and weaknesses, and offer specific feedback on a small number of areas that need improvement.
- **Clear procedures and expectations** that are clearly communicated beforehand, including the type of evidence an instructor should gather. Evaluation should be based on a shared vision of good teaching and tied to the mission and goals of the institution.
- **Peers, students and administrators receive clear guidance** on how to evaluate or how to interpret evaluations.
- **Opportunities for development**. Instructors should have access to and time for professional development.
- **Excellence is rewarded** and unsatisfactory performance is addressed in a timely manner.
- **Bias is taken into account.** All evidence and evaluators have biases, and a valid system helps account for those biases through awareness, use of multiple sources, and system design.
- **Encouraging a mastery perspective,** which focuses on improvement, risk-taking, willingness to try new techniques and persistence by adapting to evidence. This provides greater motivation than a performance orientation, which compares one instructor's performance to others and encourages competition rather than cooperation.

Improving the use of student surveys of teaching

- Students should receive the same set of instructions. Results gathered at the same time under the same circumstances are more reliable.
- Comparison of an instructor's student ratings to a mean is generally meaningless. If scores are used, they should take into account standard deviations to account for broad views of students.
- Student surveys that have the most validity contain 30 to 70 questions (Richardson, 2005; Frick et al., 2010).
- Surveys should have at least a 60% response rate to be considered valid (Richardson, 2005).

Tendencies and potential biases in student surveys

Disciplines

- Instructors in humanities and the arts generally receive the highest ratings, followed by medical sciences, natural sciences, social sciences and engineering. (Li and Benton, 2017; Murray et al., 2020).
- STEM fields tend to have lower student ratings than non-STEM fields, although many instructors in STEM fields have been slow to adopt evidence-based teaching practices.

Class type

- Small classes tend to have higher ratings than large classes (Galbraith, Merrill and Kline, 2012).
- Classes in which instructors have the most control over content tend to have higher ratings (Galbraith, Merrill and Kline, 2012).
- Quantitative classes generally receive lower ratings (Campbell, Steiner and Gerdes, 2005).
- Elective courses receive lower ratings (Campbell, Steiner and Gerdes, 2005).

Student interest, learning and expectations

- Students who have previous interest in subject matter or who are interested in a class tend to give higher ratings.
- Students who expect higher grades and those who attend more classes give higher scores on student surveys (Hamilton, 1980).
- Grades have a low positive correlation with scores given on student surveys.
- Students who perceive they have learned during a class tend to give higher ratings than students who don't, but researchers have found no correlation between actual learning and survey ratings (Clayson, 2021).
- Students put more weight on whether they like a professor than whether they learned in class (Clayson, 2021).

- Various studies have found that students are more likely to respond to student surveys when they feel engaged in a class and when a course is in their major (Young, et al., 2019).

Workloads and difficulty

- Some researchers argue that students give lower ratings to more difficult classes, but classes with higher workloads and higher difficulty tend to get *higher* scores up to a point where students consider the course too difficult and the workload too heavy.
- Time spent on tasks that students saw as meaningless leads to decreased scores (Marsh, 2007).
- Ratings are lower in courses considered too difficult or too easy.
- Students who view a class as rigorous tend to give lower ratings to instructors.
- Instructors who push students to think more rigorously may receive lower ratings (Clayson, 2009; Alauddin and Kifle, 2014).
- Students who feel that they have learned during a course and that the instructor was challenging and responsive are more likely to give higher scores (Campbell, Steiner and Gerdes, 2005).

Grades

- When students think they will receive higher grades, they are more likely to give higher scores on student surveys (Clayson, 2009).
- On course surveys, students often give an instructor the same grade they expect to receive in a course (Clayson, 2021).
- Instructors who teach classes in which students learn the most often receive middle-range scores on student surveys; those who teach classes associated with low levels of student learning often receive high or low scores (Galbraith, Merrill and Kline, 2012).
- Instructors perceived to give lower grades than students expect or more outside-class work than expected receive lower ratings (Campbell, Steiner and Gerdes, 2005).

Demographics

- Instructors who are non-native speakers tend to receive lower scores than instructors whose first language is English (Fan et al., 2019). Students who mention an instructor's accent tend to rate that instructor lower.
- Younger instructors tend to receive higher scores than older instructors.
- White teachers often receive higher scores in upper-level courses.
- Some researchers say that female students give higher scores to female instructors; other say female students give higher scores to male instructors. Other says that female students tend to give higher scores overall than male students. In some studies, female teachers receive higher scores; in other studies, male instructors receive higher scores.
- The older the student, the higher the score they generally give.

- Female instructors perceive higher impact of student ratings than male instructors do, although both female and male instructors see limited value in the surveys for improving teaching.
- Women, international students, older students and students with higher GPAs are more likely to complete student surveys (Goos and Salomons, 2017; Tucker, 2014).
- Women, faculty of color, and older faculty receive lower ratings (Campbell, Steiner and Gerdes, 2005).
- Female instructors who lecture more receive higher ratings (Campbell, Steiner and Gerdes, 2005).

Instructor characteristics

- Instructors perceived as fair, approachable, respectful or pleasant generally receive higher scores. Similarly, those with a friendly written syllabus receive higher scores.
- Instructors perceived as physically attractive sometimes get higher scores, with impact greater for male instructors.
- Some researchers have found that adjunct instructors receive higher scores than tenured and tenure-track faculty. Others, including Campbell, Steiner and Gerdes (2005), found that full professors received the highest ratings.

Misc.

- Overall, student comments tend to be more positive than negative.
- Students generally have low opinions of teaching surveys, doubting that they make any difference.
- The is little or no connection between an instructor's research and student survey ratings (Murray et al. 2020).
- There is little or no connection between student learning and student ratings (Uttl, White Gonzalez, 2017).
- Classes with earlier start times have better ratings (Campbell, Steiner and Gerdes, 2005).
- Business schools are among the heaviest users of student surveys, in part because of businesses' emphasis on consumer satisfaction (Clayson, 2009).
- Instructors who teach quantitative courses generally receive considerably lower scores on student surveys than instructors who teach qualitative courses (Uttl and Smibert 2017).
- Students often lie in student surveys of teaching, giving higher scores to instructors they like or lower scores to instructors they dislike. Students who think of themselves as customers of their college or university are more likely to provide false information (Cox, Rickard, and Lowery, 2021).

Online vs. paper

- There is little difference in scores between paper and digital formats.

- Students tend to provide more comprehensive comments in online surveys, but the proportion of positive and negatives comments remains about the same (Gakhal and Wilson, 2019).

Conclusions

Feedback from students plays an important role in effective and innovative teaching, but student surveys of teaching have played an outsized and even unfair role in the evaluation of teaching. A growing body of literature highlights those problems, and a recent ruling by an arbitrator illustrates what can happen if those problems aren't addressed. In 2018, an arbitrator at Ryerson University in Canada sided with faculty in a long-running dispute and barred the use of student survey results in determining the effectiveness of teaching, calling them "fundamentally and irreparably flawed" (Ryerson University v Ryerson Faculty Association, 2018). Organizations including the AAU, the National Academies of Science, Engineering and Medicine, the Bay View Alliance, the Research Corporation for Science Advancement, and the American Sociological Association have called on universities to make the evaluation of teaching more substantive, nuanced, and fairer. Two recent events co-sponsored by the National Academies have highlighted what has clearly become a movement among faculty and administrators to rethink university practices in evaluating teaching and in using student surveys as just one form of evidence.

The question for KU is not *whether* to gather information from students about their classes and their instructors. Rather, the university needs to consider *what types* of information students are qualified to provide and *how* the results of student survey data should be represented and used. The research cited here provides ideas and suggestions for how to do that. As this report said at the beginning, one of the few points that researchers agree on is that student surveys should be only one measure of teaching effectiveness. That is already university policy, and it is a policy that seems commonsensical even though it is often ignored in practice.

In the opening session of the National Academies event in January, Gabriela Weaver, assistant dean for student success analytics at the University of Massachusetts, Amherst, pointed to "numerous inequities, biases and barriers, baked into our traditional ways of doing things." Weaver said: "Business as usual worked until *usual* became a distant memory. Then we were all made aware of what some people have known all along: The real costs and barriers to accessing and succeeding in higher education, not only for students but also for faculty and staff, are not born equally, and they're not equitable."

We will never create a perfect student survey of teaching. We can create a fairer survey, though. We can also create a more substantive approach to evaluation that truly rewards effective and innovative teaching and that makes equity a centerpiece.

Bibliography

Abrami, P., d'Apollonia, S., and Rosenfield, S. (2007). The Dimensionality of Student Ratings of Instruction: What We Know and What We Do Not, in R.P. Perry and J.C. Smart, eds., *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, Springer, 385-456.

Alauddin, M., and Kifle, T. (2014). Does the student evaluation of teaching instrument reallymeasure instructors' teaching effectiveness? An econometric analysis of students' perceptions in economics courses. *Economic Analysis and Policy* 44, 156-168.

Alhija, F. (2017). Guest Editor Introduction to the Special Issue "Contemporary Evaluation of Teaching: Challenges and Promises." *Studies in Educational Evaluation* 54, 1-3.

Ali, A., et al. (2021). What student evaluations are not: Scholarship of Teaching and Learning using student evaluations. Journal of University Teaching and Learning Practice 18(8): 1-12.

American Sociological Association (2019). Statement on Student Evaluations of Teaching. asa_statement_on_student_evaluations_of_teaching_feb132020 (asanet.org)

Arnold, I., and Versluis, I (2019). The influence of cultural values and nationality on student evaluation of teaching. *International Journal of Education Research* 98, 13-24.

Artz, B., and Welsch, D. (2013). The Effect of Student Evaluations on Academic Success. *Education Finance and Policy* 8(1), 100-119.

Austin, A., Sorcinelli, M., and McDaniels, M. (2007). Understanding New Faculty: Background, Aspirations, Challenges and Growth, in R.P. Perry and J.C. Smart, eds., *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, Springer, 38-89.

Baker, P., and Copp, M. (1997). Gender Matters Most: The Interaction of Gendered Expectations, Feminist Course Content, and Pregnancy in Student Course Evaluations. *Teaching Sociology* 25(1), 29-43.

Baldwin, T., and Blattner, N. (2003). Guarding Against Potential Bias in Student Evaluations: What Every Faculty Member Needs to Know. *College Teaching* 51(1), 27-32.

Bartlett, A. (2005). 'She Seems Nice': Teaching Evaluations and Gender Trouble. *Feminist Teacher* 15(3), 195-202.

Benchmarks for Teaching Effectiveness Project (n.d.) Center for Teaching Excellence, University of Kansas. Benchmarks for Teaching Effectiveness Project | Center for Teaching Excellence (ku.edu)

Benton, S., and Young, S. (2018). Best Practices in the Evaluation of Teaching. IDEA Paper 69. IDEA Center. IDEA_Paper_69.pdf

Benton, S., and Cashin, W. (2011). Student Ratings of Teaching: A Summary of Research and Literature. IDA Paper 50. IDEA Center. Student Ratings of Teaching: A Summary of Research and Literature | IDEA (ideaedu.org)

Beran, T., and Rokosh, J. (2009). Instructors' Perspectives on the Utility of Student Ratings of Instruction. *Instructional Science* 37(2), 171-184.

Binderkrantz, A.S., Bisgaard, M., and Lassesen, B. (2021), Contradicting findings of gender bias in teaching evaluations: Evidence from two experiments in Denmark. Working paper funded by Independent Research Fund Denmark.

Boring, A. (2017). Gender biases in student evaluations of teaching, *Journal of Public Economics* 145, 27-41.

Boysen, G., Kelly, T., Raesly, H., and Casner, R. (2014). The (Mis)interpretation of Teaching Evaluations by College Faculty and Administrators. *Assessment and Evaluation in Higher Education*, 39(6), 641–656.

Campbell, H., Steiner, S., and Gerdes, K. (2005). Student Evaluations of Teaching: How You Teach and Who You Are. *Journal of Public Affairs Education* 11(3), 211-231.

Chavez, T.E. (2021). *Student Perceptions of Intrinsic Motivation in Completion of Student Evaluation of Teachers*. Dissertation. Grand Canyon University, Phoenix.

Clayson, D. E. (2020). *A comprehensive critique of student evaluation of teaching : Critical perspectives on validity, reliability, and impartiality*. Taylor & Francis Group.

Clayson, D. (2018). Student evaluation of teaching and matters of reliability. *Assessment & Evaluation in Higher Education*, 43(4), 666–681.

Clayson, D. (2009) Student Evaluations of Teaching: Are They Related to What Students Learn? *Journal of Marketing Education* 31(1), 16-30.

Clayson, D., and Haley, D. (2011). Are Students Telling Us the Truth? A Critical Look at the Student Evaluation of Teaching. *Marketing Education Review* 21:2 (summer), 101-112.

Coleman, M.S., Smith, T.L., and Miller, E.R. (2019). Catalysts for Achieving Sustained Improvement in the Quality of Undergraduate STEM Education. Daedalus 148(4): 29-46.

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.*
DOI: 10.1080/10528008.2021.1922924 (accessed 11 October 2021).

Croom, N. (2017). Promotion Beyond Tenure: Unpacking Racism and Sexism in the Experiences of Black Womyn Professors. *The Review of Higher Education* 40(4), 557,583.

Dennin, M., Schultz, Z.D., Feig, A., Finkelstein, N., Greenhoot, A.F., Hildreth, M., Leibovich, A.K., Martin, J. D., Moldwin, M.B., O'Dowd, D.K., Posey, L.A., Smith, T. L., & Miller, E.R. (2017). Aligning Practice to Policies: Changing the Culture to Recognize and Reward Teaching at Research Universities. *CBE Life Sciences Education*, *16*(4), es5. Also see a related report published by the AAU under the same title.

Esarey J. and Valdes N. (2020) Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education* 45(8): 1106-1120. doi:10.1080/02602938.2020.1724875

Emmelman, D. and DeCesare, M. (2007). College Students' Perceptions of Their 'Best' and 'Worst' Courses and Instructors. *International Review of Modern Sociology* 33(2): 227-244.

Fan, Y., et al. (2019) Gender and cultural bias in student evaluations: Why representation matters. *PLOS ONE* 14(2): e0209749.

Feistauer, D., and Richter, T. (2017). How Reliable Are Students' Evaluations of Teaching Quality? A Variance Components Approach. *Assessment and Evaluation in Higher Education* 42(8), 1263-1279.

Feldman, K.A. (2007). Identifying Exemplary Teachers and Teaching: Evidence From Student Ratings, in R.P. Perry and J.C. Smart, eds., *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, Springer, 93-129.

Flaherty, C. (2021). The Skinny on Teaching Evals and Bias. *Inside Higher Ed*, 17 February 2021. What's really going on with respect to bias and teaching evals? (insidehighered.com)

Flaherty, C. (2015). Flawed Evaluations. *Inside Higher Ed*, 10 June 2015. AAUP committee survey data raise questions on effectiveness of student teaching evaluations (insidehighered.com)

Fosnacht, K., Sarraf, S., Howe, E., and Peck, L. (2017). How Important Are High Response Rates for College Surveys? *The Review of Higher Education* 40(2), 245-265.

Frey, P. (1974). The Ongoing Debate: Student Evaluation of Teaching. *Change* 6(1), 47-48, 64.

Fricke, T.W., Chadha, R., Watson, C., and Zlatkovska, E. (2010). Improving course evaluations to improve instruction and complex learning in higher education. *Educational Technology Research and Development* 58, 115-136.

Gakhal, S., and Wilson, C. (2019). Is Students' Qualitative Feedback Changing, Now It Is Online? *Assessment and Evaluation in Higher Education* 44(3), 476-488.

Galbraith, C., Merrill, G., and Kline, D. (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.

Goos, M., and Salomons, A. (2017). Measuring teaching quality in higher education: Assessing selection bias in course evaluations. *Research in Higher Education*, 58(4), 341-364.

Gray, M., and Bergmann, B. (2003). Teaching Evaluations: Inaccurate, Demeaning, Misused. *Academe* 89(5), 44-46.

Groen, J., and Yves, H. (2017). The Online Evaluation of Courses: Impact on Participation Rates and Evaluation Scores. *Canadian Journal of Higher Education* 47(2): 106-120. https://files.eric.ed.gov/fulltext/EJ1154163.pdf

Hamilton, L. Grades, Class Size, and Faculty Status Predict Teaching Evaluation. *Teaching Sociology* 8(1), 47-62.

Ho, D., and Shapiro, T. (2008). Evaluating Course Evaluations: An Empirical Analysis of a Quasi-Experiment at the Stanford Law School, 2000-2007. *Journal of Legal Education* 58(3), 388-412.

Hobson, S., and Talbot, M. (2001). Understanding Student Evaluations. *College Teaching* 49(1), 26-32.

Hu, D. (2021). The Development Thread and Innovation Trend of Educational Teaching Evaluation Method. 2nd International Conference on Computers, Information Processing and Advanced Education, 310-313. https://dl.acm.org/doi/10.1145/3456887.3456956 (accessed October 2021).

Hutchings, P., Huber, M., and Ciccone, A. (2011). *The Scholarship of Teaching and Learning Reconsidered: Institutional Integration and Impact*. Stanford, California: Jossey-Bass.

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), 87-104.

Li, D., and Benton, S. (2017). The Effects of Gender and Discipline Group on Student Ratings of Instruction. IDEA Research Report 10. IDEA Center.

Lindahl, M., and Unger, M. (2010). Cruelty in Student Teaching Evaluations. *College Teaching* 58(3), 71-76.
Vereen, L., and Hill, N. (2008). African American Faculty and Student-Oriented Challenges: Transforming the Student Culture of Higher Education From Multiple Perspectives. *Journal of Thought* 43(3-4), 83-100.

Lord, T. (2008). What? I Failed? But I Paid for Those Credits! Problems of Students Evaluating Faculty. *Journal of College Science Teaching* 38(2), 72-75.

Marsh, H.W. (2007). Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness, in R.P. Perry and J.C. Smart, eds., *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, Springer, 319-384.

Martin, E. (1984). Authority in the Classroom: Sexist Stereotypes in Teaching Evaluations. *Signs* 9(3), 482-492.

Mengel, F., Sauermann, J., and Zölitz, U. (2019). Gender Bias in Teaching Evaluations. *Journal of the European Economic Association*, 17(2), 535–566.

Mitchell, K., and Martin, J. (2018). Gender Bias in Student Evaluations. *PS: Political Science& Politics, 51*(3), 648-652.

Morley, D. (2014). Assessing the Reliability of Student Evaluations of Teaching: Choosing the Right Coefficient. *Assessment and Evaluation in Higher Education* 39(2), 127-139.

Murray, D., et al. (2020). Exploring the personal and professional factors associated with student evaluations of tenure-track faculty. *PLoS ONE* 15(6): e0233515.

National Academies of Sciences, Engineering, and Medicine (2020). *Recognizing and Evaluating Teaching in Higher Education: Proceedings of a Workshop in Brief*. Washington: The National Academies Press. http://nap.edu/25685

Nowell, C., Gale, L., and Handley, B. (2010). Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. *Assessment & Evaluation in Higher Education* 35(4):463-475. doi:10.1080/02602930902862875

Onwuegbuzie, A., Daniel, L., and Collins, K. (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality and Quantity*, 43(2), 197-209.

Parker, Tara L., and Kathleen M. Neville (2019). The Influence of Racial Identity on White Students' Perceptions of African American Faculty. *The Review of Higher Education* 42(3), 879-901.

Peterson D., Biederman L., Andersen D., Ditonto, T., and Roe, K. (2019) Mitigating gender bias in student evaluations of teaching. *PLoS ONE* 14(5): e0216241.

Piña, Anthony, and Larry Bohn. Assessing Online Faculty: More Than Student Surveys and Design Rubrics. *Quarterly Review of Distance Education* 15(3), 2014, pp. 25-34.

Pittman, C. (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.

Randolph, A. (2010). What Does Racism Look Like? An Autoethnographical Examination of the Culture of Racism in Higher Education in *Tedious Journeys: Autoethnography by Women of Color in Academe*, pp. 119-148.

Richardson, J. (2005). Instruments for Obtaining Student Feedback: A Review of the Literature. *Assessment and Evaluation in Higher Education* 30(4), 387-415.

Rodriguez, J., Rodriguez, N., and Freeman, K. (2020). Student evaluations of teaching: phrenology in the 21st century?, *Race Ethnicity and Education* 23:4, 473-491.

Rodriguez, J. (2019). The Weaponization of Student Evaluations of Teaching: Bullying and the Undermining of Academic Freedom. *AAUP Journal of Academic Freedom* 10. rodriguez.pdf (aaup.org)

Rodin, M. (1973a). Can Students Evaluate Good Teaching? *Change* 5(6), 66-67, 80.

Rodin, M. (1973a). Miriam J. Rodin replies. *Change* 5(8), 7.

Ryerson University v Ryerson Faculty Association (2018). CanLII 58446 (ON LA).

Scriven, M. (1966). The Methodology of Evaluation. Publication 110, Social Science Education Consortium.

Scriven, M. (2015). Interview at Mary Lou Fulton Teachers College, Arizona State University. Michael Scriven | Mary Lou Fulton Teachers College (asu.edu)

Smith, B., and Hawkins, B. (2011). Examining Student Evaluations of Black College Faculty: Does Race Matter? *The Journal of Negro Education* 80(2), 149-162.

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.

Steiner, S., Holly, L., Gerdes, K., and Campbell, H. (2006). Evaluating Teaching: Listening to Students While Acknowledging Bias. Journal of Social Work Education 42(2), 355-376.

Stroebe W. (2016). Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluations. *Perspectives on Psychological Science* 11(6), 800-816.

Task Force on the Assessment of Teaching and Learning (2007). Faculty Senate, University of Kansas.

Titus, J. (2008). Student Ratings in a Consumerist Academy: Leveraging Pedagogical Control and Authority. *Sociological Perspectives* 51(2): 397-422.

Tucker, B. (2014). Student evaluation surveys: anonymous comments that offend or are unprofessional. *Higher Education* 68, 347-358.

Uttl B., and Smibert, D. (2017). Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ* 5:332. https://doi.org/10.7717/peerj.3299 (accessed 11 October 2021).

Uttl, B., White, C., and Gonzalez, D. (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.

Valsan, C., and Sproule, R. (2008). The Invisible Hand Behind the Student Evaluation of Teaching: The Rise of the New Managerial Elite in the Governance of Higher Education. *Journal of Economic Issues* 42(4), 939-958.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, 23(2), 191.

Walkington, L. (2017). How Far Have We Really Come? Black Women Faculty and Graduate Students' Experiences in Higher Education. *Humboldt Journal of Social Relations* 39, 51-65.

Wallace, S., Lewis, A., and Allen, M. (2019). The State of Literature on Student Evaluations of Teaching and an exploratory Analysis of Written Comments: Who Benefits Most? *College Teaching* 67(1), 1-14.

Wieman, C. (2015). A Better Way to Evaluate Undergraduate Teaching. *Change: The Magazine of Higher Learning* 47(1), 6-15.

Williams, W., and Ceci, S. (1997). ‘How Am I Doing?’ Problems With Student Ratings of Instructors and Courses. *Change* 29(5), 12-23.

Wilson, T. (1988). Student Evaluation-of-Teaching Forms: A Critical Perspective. *The Review of Higher Education* 12(1), 79-95.

Wolbring, T., and Treischl, E. (2016). Selection Bias in Students’ Evaluation of Teaching: Causes of Student Absenteeism and Its Consequences for Course Ratings and Rankings. *Research in Higher Education* 57(1), 51-71.

Wright, S., and Jenkins-Guarnieri (2012). Student Evaluations of Teaching: Combining the Meta-analyses and Demonstrating Further Evidence for Effective Use. *Assessment and Evaluation in Higher Education* 37(6), 683-699.

Young, K., Joines, J., Standish, T., and Gallagher, V. (2019). Student evaluations of teaching: the effect of faculty procedures on response rates. *Assessment & Evaluation in Higher Education* 44(1): 37-49.

Zipser, N., and Mincieli, L. (2018). Administrative and Structural Changes in Student Evaluations of Teaching and Their Effects on Overall Instructor Scores. *Assessment and Evaluation in Higher Education* 43(6), 995-1008.