Journal of Academic Ethics
https://doi.org/10.1007/s10805-024-09562-3

Why Student Ratings of Faculty Are Unethical

Daryl Close[1]

Accepted: 9 August 2024

Abstract
For decades, student ratings of university faculty have been used by administrators in high stakes faculty employment decisions such as tenure, promotion, contract renewal and reappointment, and merit pay. However, virtually no attention has been paid to the ethical questions of using ratings in employment decisions. Instead, the ratings literature is generally limited to psychometric issues such as whether a given student ratings instrument exhibits the statistical properties of reliability and validity. There is no consensus understanding of teaching effectiveness, the very attribute that students are alleged to "evaluate." What students are actually doing when they complete a ratings form—whether measuring, evaluating, reporting, judging, opining, etc.—remains unsettled in the ratings literature. If ratings are surveys of student satisfaction, they have no logical or ethical connection with teaching expertise. I argue that the administrative use of student ratings in faculty employment decisions violates basic moral principles including nonmaleficence, beneficence, professional autonomy and clinical independence, and multiple aspects of justice including due care, truthfulness, and equitable treatment. These ethical violations rule against any administrative use of student ratings in faculty employment decisions, including the "use with caution in conjunction with other evaluative methods" deployment of student ratings. My conclusion is that such use should be immediately and universally terminated. Formative use of student questionnaires as part of ordinary instructional communication and feedback between instructor and students is a separate issue and outside of the scope of this paper.

Keywords Student Ratings · Faculty Evaluation · Teaching Effectiveness · Academic Freedom · Professional Autonomy · Scholarship of Teaching

✉ Daryl Close
dclose@heidelberg.edu

[1] Department of Philosophy, Heidelberg University, 310 E. Market St., Tiffin, OH, USA

Published online: 29 August 2024

Introduction

For more than 90 years, U. S. colleges and universities have administered questionnaires to students about their professors' classes, the results of which are then inserted into the institution's faculty employment processes. The questionnaires, known variously as "student ratings," "student evaluations of teaching effectiveness," "course evaluations," "student evaluations of teaching (SETs)," and "student evaluations of faculty (SEFs)" solicit responses from students, typically on some sort of quantitative Likert response format, with or without written comments. Those student responses are then treated as evaluations, measurements, or reports regarding the professor's "teaching effectiveness," "quality of teaching," "excellence in teaching," and the like.

Administrative decisions about tenure, promotion, faculty awards, dismissal, contract renewal, salary increases, and merit pay are based in part—sometimes entirely—on these purported evaluations or measures. Following common usage, I will refer to these questionnaires generically as "student ratings" in what follows, even though we cannot assume without argument that students are actually rating, evaluating, or measuring the clinical skills[1] of their professors. I will also adopt—without argument—the common expression, "teaching effectiveness," even though the sense of causality embedded in the expression is itself questionable and is not present in expressions such as "excellence in teaching" or "quality of teaching."[2]

This paper does not survey any plausible arguments to morally justify the administrative use of student ratings in faculty employment decisions simply because there are none to be found. Franklin and Theall (1989) briefly discuss statistical validity and reliability as necessary conditions of "acceptable use," but do not appear to be using the term as an ellipsis for "ethically acceptable use." I found no evidence in the ratings literature of subjecting the administrative use of student ratings to ethical standards such as those in National Council on Measurement in Education's (NCME) *Code of Professional Responsibilities in Educational Measurement* (2016) or the American Educational Research Association, et al. (AERA) *Standards for Educational and Psychological Testing* (2014).

While there has been no general treatment of the ethical violations in using ratings in employment decisions, some ratings researchers have stated their opposition to this use of ratings on the grounds of gender bias, e.g., see Boring et al., 2016. The logical independence of the moral justification of a practice from the quality of statistical reasoning involved in that practice is treated in Sect. Student Ratings Are Not a Single Thing.

The use of student ratings in university faculty employment decisions is studded with ethical problems. There are prominent violations of basic moral principles[3] including non-

[1] I use the term, "clinical skills," to indicate those skills in the professions that characterize a best practice professional-client relationship. While clinical skills are usefully distinguished from the professional's discipline-specific knowledge, they are inherently dependent on that knowledge.

[2] The following issues are important but beyond the scope of this paper: the theoretical question of whether the expression "teaching effectiveness" denotes an actual phenomenon; the logically separate issue of student questionnaires used by faculty as part of ordinary instructional communication between instructor and students; and the details of peer review as the gold standard in the traditional professions for evaluation of practitioners.

[3] See the Beauchamp et al. (2008) set of normative principles widely adopted in the professional ethics literature.

maleficence, beneficence, professional autonomy and clinical independence, and multiple aspects of justice including due care, truthfulness, and equitable treatment.

Such principles are hardly unknown to educational researchers. For example, the NCME *Code of Professional Responsibilities in Educational Measurement* refers to "honesty, integrity, due care, and fairness," and to the right to privacy, concepts that are grounded in the principle of justice. The bible of professional standards in testing is the AERA *Standards for Educational and Psychological Testing* (2014). The Standards document defines "tests" very broadly, including surveys, credentialing observations, and professional assessments (see pp. 174–175). Fairness in testing and avoiding harm in testing are discussed. The AERA Standards also refers testers to professional ethics codes. Here, the NCME Code would certainly apply.

Instead, the ratings literature is almost entirely limited to psychometric issues such as whether a student ratings instrument at a given institution exhibits the statistical properties of reliability and validity. It is striking that after many decades of use, virtually no attention has been paid to the institutional ethical questions of using ratings in high stakes faculty employment decisions such as tenure, promotion, contract renewal and reappointment, and merit pay. This paper seeks to address that lack of attention. I will identify ethical problems that rule against any administrative use of student ratings in faculty employment decisions. This includes the "use ratings with caution in conjunction with other evaluative methods" defense of student ratings. My overall conclusion is that all use in employment decisions should be immediately and universally terminated.

My conclusion is neither novel nor extreme. For example, large institutions such as the University of Southern California have already eliminated the use of student ratings in faculty employment decisions (Flaherty, 2018). A recommendation arising from a 2014 survey of faculty (9,314 respondents) conducted by the American Association of University Professors (AAUP) Committee on Teaching, Research, and Publication, stated that "institutions should evaluate teaching as seriously as research and scholarship" where "faculty members within departments and colleges—not administrators—should develop instruments and determine practices (peer review, classroom visits, teaching portfolios)" (Vasey & Carroll, 2016).

This paper presents three arguments that independently support my overall conclusion that using ratings in faculty employment decisions is unethical and should be ended. The first argument concerns the lack of agreement among ratings researchers and ratings users about the very nature of ratings. For example, concepts such as "student evaluations," "student ratings," and worse, "teaching effectiveness," have no consensus definitions. These claims are defended in Sect. Justice and Truth: What Are Student Ratings, and subsections Student Ratings Are Not a Single Thing, There is No Consensus Definition of Teaching Effectiveness, Are Students Evaluating, Judging, or Measuring Teaching? and The Fallacy of the Student Competency "Myth" Argument. The second argument concerns the serious harms of a wrongful nonrenewal of a teaching contract, a wrongful denial of merit pay, or a wrongful denial of tenure or promotion. It follows that using administrative authority to induce students to participate in a process that may lead to such harm is morally wrong. Demonstrable ratings biases such as age, gender, race, time of day, and course subject matter, reveal unjust student prejudices but nothing about the professor's professional competence. These claims are defended in Sect. Ratings Harm. Third, if student ratings of teaching are actually opinion polls of student satisfaction, then no factual conclusions can be drawn

about either the clinical skills or disciplinary knowledge of the practitioner. This claim and related violations are defended in Sect. Learning, Customer Satisfaction, and Academic Freedom.

Justice and Truth: What Are Student Ratings?

Examining the precise nature of student ratings might appear to be a fool's errand since everyone already knows what ratings are. Such an assumption would be a mistake. We actually know little or nothing about student ratings at the foundational level.

Student Ratings Are Not a Single Thing

Student ratings are reported as involving multiple equity biases (Kreitzer & Sweet-Cushman, 2022). So, it is not surprising to hear arguments in faculty discussions such as "Student ratings have been shown to be gender biased and therefore ratings here at our university are invalid and ethically questionable" or conversely, "Student ratings have been established as a valid assessment of teaching effectiveness; therefore, ratings here at our university are valid and ethically unobjectionable." Both of these arguments fail, but for two different reasons.

First, the moral acceptability of a faculty evaluation system is logically independent of the statistical validity or reliability of the system. One can construct a long list of counterexamples to the proposition that the statistical validity or reliability of some human or animal study is sufficient for the moral permissibility of that study. For example, the tragic ethical faults of the clinical study, "Tuskegee Study of Untreated Syphilis in the Negro Male" (Kampmeier, 1972), are not a consequence of statistical weaknesses. One could assume that the study was statistically without flaw and the deeply unethical nature of the study would remain. So, statistical validity is simply not *sufficient* for moral acceptability.

From the converse direction, the idea that statistical validity is a *necessary* condition for the moral permissibility of a faculty evaluation system, or any human subject research is also false. If it were true, we would have to reject a substantial fraction of, say, surgical procedure research as being morally wrong because of small sample size, the lack of adequate controls, and inherent nonrepeatability. Consequently, the apparently unending validity debate in the ratings literature is a distraction from the ethical problems of using ratings in employment decisions. This logical separation of moral permissibility from statistical strength of reasoning has not escaped attention in the ratings literature, e.g., "Unbiased, Reliable, and Valid Student Evaluations Can Still be Unfair" (Esarey & Valdes, 2020). Of course, it was the unethical nature of the Tuskegee study that led the U. S. Congress to pass the National Research Act (1974), leading to the creation of Institutional Review Boards, precisely because the ethical dimensions of human subject research lie outside the scope of the statistical properties of the research.

The second and deeper error is that there simply is no species known as "student ratings." The ratings validity debate has distracted our attention from the fact that the expression, "student ratings" doesn't denote a specific thing. As Scriven (1995) points out, ratings instruments vary widely from one study to another. Examples abound. There are several different, sometimes extensively researched ratings forms such as.

- Edwin Guthrie's 1925 instructor ranking experiment at the University of Washington (Guthrie, 1927),
- Hermann Remmers' and George Brandenburg's seminal *Purdue Rating Scale for Instructors* (Brandenburg & Remmers, 1927, 1928),
- William Wilson's 1929 first-time, faculty-wide administration of a "rating blank" at the University of Washington (1932),
- Herbert Marsh's (1982) *Students' Evaluations of Educational Quality* (SEEQ),
- Kansas State University's *IDEA Student Ratings of Instruction* (Anthology, 2022; Campus Labs, 2020),
- Michigan State University's *Student Instructional Rating System* (SIRS) (Michigan State University Board of Trustees, 2011),
- the University of Illinois at Urbana-Champaign's *Instructor and Course Evaluation System (ICES)* (University of Illinois at Urbana-Champaign, 2022), and many others.

Additionally, there are a large number of differing, homegrown instruments used at colleges and universities.

Despite this extensive and very visible diversity of student ratings instruments, the expressions "student ratings," "student evaluations of teaching effectiveness," "course evaluations," and other variations continue to be used by ratings researchers, faculty, and administrators as though there were a clear and universally settled singular reference of these terms. In reality, there is no singular reference. It is therefore a serious conceptual error to treat student ratings as a single thing with regard to statistical analysis.

This conflation of research conclusions from different ratings instruments is pervasive in the research literature. We can find examples of this misuse of language in the very publication titles themselves. For example, an entire issue of *New Directions for Institutional Research* is titled, "The Student Ratings Debate: Are They Valid? How Can We Best Use Them?" (Theall, Abrami, & Mets, 2001) when in fact, there is no "they" or "them." A paper title—one of many—that exhibits the same error is "On the Validity of Student Evaluation of Teaching: The State of the Art" (Spooren et al., 2013).

There is No Consensus Definition of Teaching Effectiveness

The proponents of student ratings may say, yes, we know that there are many different instruments for student ratings, but why is that of any concern? After all, they are all measuring teaching effectiveness, just with different tools, like measuring the temperature of a liquid with different sorts of thermometers. But unlike thermometers, where there is a universal definition of heat that governs the design of thermometers, there is *no consensus definition of teaching effectiveness*, the very trait, or collection of traits, that students are alleged to evaluate, measure, or report when they complete a ratings form. Defining a theoretical concept is often a complex problem in the social sciences, of course, but using ratings in employment decisions as if there were even an informal agreement about teaching effectiveness is ethically unacceptable in the face of a long history of no agreement.

The alarming absence of a definition has been observed routinely in the higher education research literature for more than 80 years. Lily Detchen's, 1940 paper, "Shall the Student Rate the Professor?" is an early example of awareness of the problem:

> [I]n institutions where students have been asked to rank the qualities they considered necessary for successful teaching, there has been found within the institution close agreement on the relative values of teaching traits, but there is a fickle variation in the requisite attributes as they are described from institution to institution (2 [Bowden 1926]; 4 [Cattell 1931]; 5 [Champlin 1931]; 7 [Clinton 1930]; 8 [Flory 1930]; 12 [MacDonald 1931]; 13 [Mills 1931]). (Detchen, 1940, p. 149)

Here are further examples of this concession in chronological order made by leading ratings researchers, all observing the failure to reach even an approximate definition of teaching effectiveness:

- It must be admitted that we shall never reach a completely factual basis for evaluating the operation of teaching. (Guthrie, 1954, pp. 1–2)
- It seems most unlikely that any one set of characteristics will apply with equal force to teaching of all kinds of material to all kinds of students under all kinds of circumstance…(Doyle, 1983, p. 27)
- [S]tudent ratings forms, each purported to measure instructional effectiveness, were not consistent in their operational definitions of instructional effectiveness. Thus, no one rating form represents effective instruction across contexts. (d'Apollonia & Abrami, 1997, p. 1199)
- What construct domain do student rating items attempt to represent? Is there a universal set of characteristics of effective teachers and courses that should be used as a target? Unfortunately, no such set appears to exist. (Ory & Ryan, 2001, pp. 31–32)
- SET researchers agree that SET and SET instruments should capture multiple aspects (dimensions) of good teaching practice. Due to the absence of an agreement with respect to the number and the nature of these dimensions, which should be based on both the theory and empirical testing, SET instruments vary greatly in both the content and the number of dimensions. (Spooren, Brockx, & Mortelmann, 2013, p. 607)
- The reliability and validity of rating forms have also received a great deal of attention in the literature with mixed results. This literature is complex in part because there is no single agreed upon definition of the construct of teaching effectiveness…This lack of agreement has contributed to the development of numerous student evaluation of teaching scales (SETs) with varying numbers of items and variable item content. (Shook & Greer, 2015, p. 89)

It is difficult to exaggerate this failing. It constitutes both a logical and an ethical problem. One immediate conclusion is that generalized claims of the validity and reliability of student ratings should be avoided purely on the grounds of scientific truthfulness. Instead, there should be highly qualified conclusions about specific definitions of teaching effectiveness and the specific ratings instruments that allege to measure the existence or degree of teaching effectiveness, as defined by that specific instrument.

Second, since cross-form comparisons cannot be logically made in the absence of a common definition, the ethical problem of justice becomes critical. This is because what students are actually evaluating or measuring—if they are indeed doing either—is sui generis and should be better named with concept-neutral terms that are individually paired with

each different ratings instrument, e.g., "Alpha," "Beta," "Gamma," "Delta," etc. This would prohibit fallacious cross-form statistical comparisons so common in the ratings literature.

This is not a situation of "Oh, well, we all basically agree on the definition." We have just seen that that claim is facially false. Worse, the depth of disagreement is profound. Consider the characteristic of *clarity of presentation*. For example, Herbert Remmers' (1929) pioneering *Purdue Rating Scale for Instructors* contained 10 items, including "Interest in Subject," "Fairness in Grading," "Personal Appearance," and "Presentation of Subject Matter." This last item is scaled, with response choices ranging from "Indefinite, involved, and monotonous" to "*Clear*, definite, and forceful" [emphasis added]. Clarity of presentation subsequently appears on other rating scales such as Item 199 on the University of Michigan's *E&E Teaching Questionnaires*: "The instructor explained material *clearly* and understandably" [emphasis added] (reproduced in Gravestock & Gregor-Greenleaf, 2008, p. 103), and Item 10 in the "Teaching Methods" section of the *IDEA Diagnostic Feedback* form: "Explained course material *clearly* and concisely" [emphasis added] (Li et al., 2016, p. 43).

Surely, everyone can agree on the "clarity" dimension of effective teaching, correct? No. The instructional characteristic of clarity of presentation has been specifically challenged as a ratings-form item by Mason Marshall and Aaron Clark in their paper, "Is Clarity Essential to Good Teaching?" (2010). Marshall and Clark, citing a teacher no less than Socrates as an example, argue that deliberate vagueness and lack of clarity can be important pedagogical techniques. Such a fundamental disagreement suggests that characteristics of "effective teaching" vary from discipline to discipline and even from instructor to instructor, rendering university teaching, like medicine, as much an art as a science.

There is another reason that a standard definition of teaching effectiveness has not emerged over decades of research and analysis: some researchers have asked *students* to define "teaching effectiveness" where the students are the very subjects of the study who then measure, report, evaluate, or opine about teaching effectiveness (see Detchen, 1940; Guthrie, 1927; Marsh, 1984; University of Chicago, 1926; Wotruba & Wright, 1975).

Wotruba and Wright (1975, p. 654) give a specifically political reason for this logically odd procedure of involving students (clients) in defining best clinical practices of a professional: "[we should] include the concerned parties sufficiently in the development of the [teaching evaluation] instrument so that they will be more open to accepting the results." This practice is unheard of in other traditional professions such as medicine or law, thus challenging the very idea of university teaching as a profession.

A third reason that there is no consensus on teaching effectiveness is that properly- worded ratings items often must be worded to request student opinion rather than student judgment, which in turn moots the question of item validity with regard to quality of teaching. One of the founders of student ratings, Wilson (1932), discusses the University of Washington rating item, "To what extent has this course been interesting to you?" He observes this wording must be favored over "How interesting was the course?" because

> If all the members of a class say that the course was interesting to them, it would be absurd to ask, "But was the course really interesting? Might not the course actually have been dull, and the students have been mistaken in thinking that it was interesting?"…If the students report that the course is interesting and the visitor reports that it is dull, the only conclusion that can be drawn is that the course is interesting to the

> students and dull to the mature visitor. If either set of appraisals is taken as a criterion, the other set is invalid. (Wilson, 1932, pp. 79–80)

Wilson's point in this passage is to show that student opinions are irrelevant to the question of validity when treated as correspondence with expert peer opinion. The "interesting to you" criterion thus emerges as a controversial component of teaching effectiveness precisely because it can't serve the goal of test validity. Stark (2016) makes a similar observation in his distinction between measuring student experience ("Did the student enjoy the class?") as opposed to a measurement of student judgment ("Was the instructor fair?") or student memory ("Accurately report[ing] the number of hours per week they typically spend working on a course."). As will be seen below, this problem feeds the controversy over what students are doing when they fill out a rating form. Researchers like Marsh (1984, p. 725) may simply choose one horn of the dilemma and deny the legitimacy of any peer evaluation of faculty teaching, thus leading to a complete reliance on student ratings for the evaluation of teaching expertise.

So, in the 80-plus year interval between 1940 and the present, ratings researchers have demonstrably failed to reach a consensus definition of teaching effectiveness, despite the essential—and ethically deal-breaking—need for such a definition in student ratings. Wachtel (1998) cites the lack of a definition of effective teaching as one of the reasons for "faculty hostility and cynicism" towards student ratings. Such entirely reasonable faculty skepticism is now itself a subject of investigation among ratings researchers under the label "myths of student ratings" (e.g., Cohen, 1990) in which faculty criticisms of student ratings are treated as having no foundation and arising out of ignorance.

The obvious ethical conclusion is that this lack of a consensus definition of teaching effectiveness in student ratings is a sufficient reason to immediately terminate all use of student ratings in faculty employment decisions. This conclusion follows easily from the principle of nonmaleficence and the principle of justice, especially truthfulness, due care, and fairness. We simply do not know from instructor to instructor, course to course, campus to campus, what is being rated, evaluated, or opined about. The stakes for the ethical treatment of university faculty are simply too high to tolerate such deep and persistent ignorance.

It is important to note that this conclusion does not depend on some radical premise that psychology is not a science. The point is simply that it is unethical to use central, but unsettled theoretical constructs as if they were settled when using them to make faculty employment decisions.

Are Students Evaluating, Judging, or Measuring Teaching?

There is a long-standing disagreement about what students are actually doing when they complete a ratings form. On the one hand, the use of the terms, "rating" and "evaluation" to describe what students are doing when they respond to ratings items can be traced from the earliest student ratings research in the 1920s and 1930s, e.g., Edwin Guthrie at the University of Washington and Hermann Remmers at Purdue University. Guthrie speaks of "student judgments of teachers" (1927, p. 175), while Remmers speaks of both "student judgments" about instructors and student "evaluation" of instructors (1933, p. 22). The *Manual for Purdue Rating Scale for Instructors* states that "[teacher] traits must be of such nature that they are fairly susceptible to student observation and judgment (Brandenburg & Remmers, 1928,

p. 31). Much later, we find Herbert Marsh's (1982) well-known "Students' Evaluations of Educational Quality" (SEEQ) ratings form. The acronyms, "SET" (student evaluation of teaching) and "SEF" (student evaluations of faculty), are by then in widespread use. For example, an EBSCOHost literature search shows over 1,000 articles with titles containing the search string, "student evaluation of teaching."

Edwin Guthrie's, 1927 research is regarded as the first formal study of student ratings, and the title Guthrie chooses for his report is instructive: "Measuring Student Opinion of Teachers" (1927). For Guthrie, the students' opinions obviously are judgments and the point of collecting those judgments is for faculty employment decisions. The opening paragraph of the paper begins with a question that quickly emerges as purely rhetorical:

> Are college students competent judges of the quality of teaching in their courses? Quality of teaching must be and is judged as a basis for promotion and pay. The judgment of it is usually made by persons who know it only by hearsay, through faculty and student gossip, or, occasionally, by effects evident in other classes. Students have an opportunity for observing the quality of teaching that no fellow-teacher, head of department or school authority ever enjoys. They alone have a direct classroom equivalence with their teachers. (Guthrie, 1927, p. 175)

In this two-page paper, Guthrie establishes a now common argument for student competence in evaluating faculty that persists to this day, viz., that students' direct exposure to the professor uniquely qualifies them as evaluators of the professor's clinical skills. As I will argue below, this "long exposure" defense of student competence is defective, but despite that, it continues to be the first-line defense of the use of student ratings in faculty employment decisions.

Like Guthrie, Remmers, Marsh, and others, contemporary researchers Michael Theall and Jennifer Franklin describe student raters as providing "opinions or estimates," "the value they place on their experiences," and "summary opinions," e.g., comparing instructor performance (1990, p. 1). But, Theall and Franklin challenge Guthrie's view of student raters as evaluators of teaching quality. Their answer to the question, "Are students actually the evaluators?" is "No." Their premise for this conclusion is that the "student's role" doesn't include "making a decision about merit or worth," which they assert to be a standard component of evaluation. Theall and Franklin thus appear to draw a distinction between evaluating and judging. However, whether students can form judgments about their professors but not be regarded as evaluating them seems to be a distinction without an ethical difference.

Linse (2017) also rejects the common "student evaluation" terminology, asserting that "[s]tudent ratings are not faculty evaluations" and that "ratings researchers are clear to differentiate between the producers of the data (students) and the users of the data (faculty and administrators) for both improvement and evaluative purposes" (p. 2). However, Linse is just factually wrong about this linguistic practice because many—perhaps the majority of—student ratings researchers regard student raters as *evaluating* faculty and typically do not draw a distinction between opining and evaluating. If Linse is prescribing what should be the case, viz., that student ratings should not be treated as evaluations of faculty, then the

only conclusion is that historically, neither most ratings researchers, nor faculty, nor administrative users of student ratings have generally followed that prescription.[4]

So, the concepts of "opinion," "judgment," "evaluation," and related concepts are not uniformly treated in the ratings literature. Guthrie thinks that student opinions about teaching are judgments, though possibly poor ones if reliability and validity cannot be determined (1927, p. 175), while Theall and Franklin, and Linse treat the concept of "opinion" as excluding evaluation. Linse promotes the distinction between expressing an opinion and evaluating via the idea that students are merely producing "data" that are then evaluated by others. But whether the data are about students' own mental states or instead are about external phenomena is decisive. If the latter, then data accuracy arises. If the former, evaluation, per se, doesn't arise at any level. This is a central topic in Sect. Learning, Customer Satisfaction, and Academic Freedom below.

The problem with the term "opinion" is that it is ambiguous in ordinary language. Sometimes, an opinion simply is a judgment. For example, legal opinions are literally written judgments. When a patient seeks a medical opinion from a physician, the patient expects a medical evaluation or judgment. If you ask a professional engineer to perform an inspection of a home that you want to buy, the engineer will produce an opinion about the condition of the home, i.e., an expert judgment or evaluation of the home's structural properties. On the other hand, "opinion" in the sense of "opinion poll" is a preference or an expression of personal taste, e.g., preferring one flavor of ice cream over another. The reason that we see differing uses of "opinion" in the ratings literature arises from this ambiguity of "opinion."

The concept of rating is also problematic. We might agree with Theall and Franklin that students should not be regarded as evaluating faculty, but Theall and Franklin have no concerns with the concept of rating, as in "student raters" (1990, p. 2). This would appear to be a mistake. A dictionary definition of rating is to estimate the value of something. Music students enter contests in which their performances are rated, i.e., evaluated, on the basis of various criteria. A faculty search committee rates the applicants, i.e., evaluates them, and often places them in rank order.

Finally, the concept of observation is unclear. Although Guthrie speaks of students "observing the quality of teaching," observation, per se, is not an evaluative judgment in either ordinary language or evaluation theory.[5] For example, it could correctly be said of me that I observed a rare species of bird for my locale, but that at the time, I did not know or believe that I did because I thought that it was a house sparrow that is common in my location. Regardless, for Guthrie and others of like mind, students must be both observing teaching and subsequently forming an evaluative judgment about the quality of teaching which they then exhibit in their responses to the questionnaire prompts.

Consequently, we have yet another reason for ending the use of student ratings in faculty employment decisions: there is wide disagreement about what students are actually doing when they fill out a ratings form.

[4] The faculty at the author's institution has replaced the expression "student evaluation" and cognates with "student questionnaire" throughout the faculty manual and other student feedback policy documents.

[5] There is a long-standing debate in the philosophy of perception as to whether visual perception is fundamentally non-epistemic, e.g., that seeing does not entail any belief or judgment about what is seen. For example, see Close (1976), Dretske (1969), and Warnock (1965). We should assume that Guthrie means "consciously noticing" when he speaks of students "observing" teaching.

The Fallacy of the Student Competency "Myth" Argument

We first encountered the assertion that students are qualified to evaluate their professors in Guthrie, viz., "Students have an opportunity for observing the quality of teaching that no fellow-teacher, head of department or school authority ever enjoys. They alone have a direct classroom equivalence with their teachers" (1927, p. 175). Versions of this claim are found from the earliest days of rating research (see Aleamoni, 1987, p. 4; Detchen, 1940, p. 147; Guthrie, 1927, p. 175; Kulik & McKeachie, 1975, pp. 210–211; Remmers, 1958, p. 20; Theall & Franklin, 2001, p. 48). Faculty who disagree are argued to be captives of a "myth."[6] Here, we are concerned only with the so-called myth that "Students are not qualified to make judgments about teaching competence" (Cohen, 1990, p. 124).

First, we have already established that the expression, "student ratings," does not refer to a single thing predicated on a consensus definition of teaching effectiveness. Therefore, attempting to create a distinction between "facts" and "myths" regarding ratings, per se, is a nonstarter. Moreover, if student ratings are not evaluations, judgments, or observations of teaching expertise, but simply personal opinions or expressions of satisfaction, then the entire question of student competence is moot. One can't have it both ways, i.e., arguing that students are competent evaluators of faculty and at the same time asserting that they are not evaluating faculty when they complete a ratings form, but only expressing personal beliefs.

There is a further error here. The standard defense of student competency in evaluating faculty typically cites the lengthy exposure of students to teaching. I call this the "long exposure" argument for students as competent evaluators of university teaching. The idea that long, direct observation of some phenomenon is sufficient to bestow either evaluative or reportorial expertise on the observer is patently false.

For example, just because I spend a great deal of time, year after year, looking at the show chickens in the poultry barn at the county fair, does not thereby mean that I am qualified to report on the relevant qualities of the chickens, judge them, rate them, rank them, or otherwise evaluate the chickens. More pointedly, just because 4-H member Mary has spent far more time with her show chicken than the poultry judges, does not mean that she is better qualified than the judges to evaluate, or report on, her chicken. Even less defensible would be the claim that Mary "alone" (Guthrie, 1927, p. 175) can judge her chicken or report on her chicken's relevant qualities, or that 4-H'ers like Mary "are pretty much the only ones" (Remmers, 1958, p. 20) who can judge their chickens. The long-exposure defense of student evaluation/reportorial competence[7] is such an obvious error that it is more than a little surprising to see how persistent it has been in the ratings literature.

[6] The term "myth" in this context appears to have been introduced by Cohen (1990, p. 123). The subsequent ratings literature is dotted with allegations that critics of ratings subscribe to a variety of so-called "myths." For example, see Aleamoni (1999), Cohen (1990), ICES (2023), and Theall (2003).

[7] Whether students are qualified *reporters*, per se, is an empirical question that is routinely conflated with the question of whether students are competent judges. Unfortunately, there is little research on student reportorial skill to be found. Helpful studies might include student reports of an instructor returning graded work in a timely fashion, *as verified by a disinterested observer*, reports of the instructor's speech volume, *as verified by random sampling with a sound level meter*, or student reports of unfair grading practices *as verified by a disinterested instructional expert*. Such empirical studies would be interesting to a senior colleague of impeccable integrity who took great pride in returning any graded work at the very next class meeting. It was a regular irritation to him when students routinely gave him middling Likert scores on the promptness of his return of graded work.

A report by the International Council on the Future of the University correctly separates long exposure and evaluative competence like this:

> Persons who are outside the class-room cannot see or pass judgement on the teacher's fulfillment of his obligations to his students. Only continuous presence and adequate knowledge permit that and the students, who are the only ones continuously present, do not have adequate knowledge…The matter comes down therefore to each teacher's own sense of obligation and his voluntary submission to it. (Shils, 1983, p. 47)

Shils could not have better expressed the essence of professional integrity, and therefore, how the administrative use of student ratings is a fundamental intrusion on the professional autonomy of the university professor (e.g., see Hashimoto, 2006). On the most minimal grounds, students are not competent to evaluate or report on the teaching expertise of their professors, and so we have yet another argument that it is unethical to use student ratings in faculty employment decisions.

Why is it, then, that whenever we academics need to make life-changing decisions about our colleagues, we turn to persons (our undergraduate students) for whom we have *no* empirical evidence of knowledge or skills in evaluating the pedagogical skills—clinical skills in the general professional sense—of faculty? If the practitioners of a profession are unsettled about the nature of a given professional trait, how can we imagine that our students have knowledge of that trait that is sufficient to provide context for responding to items on a rating form? It would be question-begging to argue that they *must* have such knowledge because we have student opinion surveys that allege to support a theoretical construct named "teaching effectiveness" when the very definition of that construct is demonstrably unsettled.

Ratings Harm

There is no debate that wrongful faculty employment decisions constitute serious harms. It follows that it is morally wrong to use administrative authority to induce students to participate in a process that may lead to such harm. The basic ethical principles here are justice, beneficence, and respect for persons referenced in the *Belmont Report* (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). These principles are very familiar to medical researchers, psychologists, and other human subject researchers.

In his book, *Evaluating Teaching*, Kenneth Doyle expresses concern with the potential for harm to faculty in the use of student ratings for employment decisions. Doyle says that

> …some purposes for evaluating teaching require information of higher quality than do other purposes. A reasonable ethic in this regard would be that the greater the potential for harm to individuals, the more rigorous the information needs to be… evaluations for course diagnosis and improvement can proceed with information of less rigor than would be required for personnel decisions, in which considerable harm to individual faculty can occur. (Doyle, 1983, p. 16)

The idea that administrative use of student ratings can cause harm to faculty is also echoed in a recent paper that reports that course subject matter is "strongly associated with SET ratings." The authors determined that

> [p]rofessors teaching quantitative courses are far more likely not to receive tenure, promotion, and/or merit pay when their performance is evaluated against common standards. Moreover, they are unlikely to receive teaching awards. (Uttl & Smibert, 2017, p. 11)

Since it is prima facie highly improbable that the tens of thousands of quantitative courses would be taught primarily by poor instructors in comparison with instructors of nonquantitative courses, we can be confident that such ratings-based denials of tenure, promotion, and merit pay are wrongful and unjust harms.

Another source of ratings harm to faculty is the common practice of anonymous student ratings. Anonymity can result in psychological phenomena such as deindividuation and moral disengagement. Deindividuation and moral disengagement in student ratings are addressed by Lindal and Unger in "Cruelty in Student Teaching Evaluations" (2010). The authors observe that the widely-defended practice of student anonymity in ratings procedures diminishes moral restraints. This effect

> results in a lowered threshold for the expression of usually unacceptable and unexpressed behaviors (Zimbardo, 1969; Deiner et al., 1976; Rogers & Ketchen, 1979)… The structure of the collection process itself, involving a group situation, heightened emotional arousal, and anonymity, encourages deindividuation and may allow the mechanisms of moral disengagement to operate, permitting behavior that students would never engage in face-to-face. (Lindahl & Unger, 2010, p. 73)

Due care requires protecting faculty from false negative ratings, including libelous anonymous written comments, where the faculty member has no due process to confront the accuser, demand evidence, or challenge defamatory claims made by the student. As noted above, there is considerable moral and legal exposure of the institution itself when it uses its authority to encourage students to engage in potentially harmful conduct. Think "Milgram experiments."[8]

Third, the serious injustice of student ratings being biased with regard to instructor characteristics such as gender, race, ethnicity, religion, accent, or physical attractiveness is obvious.[9] The existence of bias in student ratings instruments has long been a subject of investigation. Wachtel (1998) provides a useful historical record of student ratings bias studies from the 1920s through the mid-1990s. Gender bias, especially, has received considerable attention for decades (see Boring et al., 2016; Mitchell & Martin, 2018). However, as Merritt (2008) observes, research into racial bias in student ratings is not as extensively researched as gender bias.

[8] Stanley Milgram's "electric shock" experiments at Yale in the early 1960s concerned obedience to authority. Unlike student ratings, there was no actual or potential harm to anyone, only the appearance of harm. See Milgram's original paper (1963) and his discussion in the popular press (1973).

[9] The injustice of bias in faculty employment decisions is logically independent of the impact of biased samples on statistical validity.

How should we respond to the issue of equity bias? Commercial ratings instruments such as the *IDEA Student Ratings of Instruction* form generate department-level and campus-level statistical data for comparative use (see Campus Labs, 2020),[10] which encourages inequitable instructor comparisons. Consequently, researchers such as Kreitzer and Sweet-Cushman (2022) argue against using ratings for cross-faculty comparisons:

> Because one way that equity bias manifests is through lower evaluations for astereotypic instructors (i.e., women in male-dominated fields and vice versa), comparisons across faculty members further disadvantage already marginalized faculty. (p. 78)

Institutions such as the University of Southern California are reported to have ended the use of student ratings in faculty employment decisions on the grounds of equity bias alone (Flaherty, 2018). In opposition to this, Kreitzer and Sweet-Cushman advise us that there is no need to "throw the baby out with the bathwater" (p. 78). Instead, they say, ratings "should be properly contextualized and used with caution" (p. 78). Rowan et al. (2017), for another deployment of the baby-bathwater defense of student ratings. This common defense fails because even if equity biases were statistically "contextualized" away, we have already demonstrated that any use of student ratings in faculty employment decisions is morally indefensible for multiple reasons other than equity bias. (Note that those disqualifications of using ratings in employment decisions have nothing to do with faculty using student questionnaires for ordinary instructor-student communication and feedback.)

Finally, there is a serious harm-related problem with ratings regarding institutional research board (IRB) review. The routine understanding of U. S. law is that student ratings are exempt from IRB review (Protection of Human Subjects, 2009, Sect. 46.101; also see Office of Human Research Protections, 2017). The legal exemption of student ratings has been challenged on several grounds (Sullivan, 2011). Of particular concern is the Federal definition of research as "a systematic investigation, including research development, testing, and evaluation designed to develop or contribute to generalizable knowledge" (Protection of Human Subjects, 2009, Sect. 46.101).

For example, the use of a commercial ratings form in which the results of an institution's student ratings are merged with other institutions' data and then summary statistical data are distributed to all participating institutions appears to "contribute to generalizable knowledge." One researcher reports that a substantial fraction of respondents in an evaluation study were found to "use their findings to share best practices and lessons-learned. This type of dissemination is typically considered a contribution to generalizable knowledge…" (Donovan, 2013). Regardless of the presumptive lawful exemption of student ratings from IRB review[11], the exemption still presents serious ethical exposure at both the individual

[10] I use the term, "commercial," loosely. IDEA is a not-for-profit company whose course evaluation products are sold by Campus Labs, which in turn is now owned by Anthology, a for-profit, privately held company, majority owned by Veritas Capital at the time of this writing.

[11] Whether failing to submit student ratings for IRB review involves legal exposure depends in part on student status. For example, exemption from IRB review requires excluding certain types of student status, e.g., being a prisoner, being a minor, or being pregnant, all conditions routinely encountered in college classrooms. Institutions that allow minors to take undergraduate courses may alone compel IRB review of any ratings instruments involving such students. It doesn't help that the U. S. Department of Health and Human Services IRB exemption decision tree (2020) employs terms the meanings of which are either vague or unsettled.

and institutional level. From a moral perspective, IRB review of ratings forms is essential because at the very least there is the possibility of social coercion.[12]

In summary, despite the many years of research regarding equity bias in student ratings, student moral disengagement, social coercion, unjustified comparisons of instructors, and lack of IRB review of ratings, there seems to be little appetite for drawing the unavoidable conclusion that institutions should immediately stop using student ratings in faculty employment decisions. The principles of justice and due care to avoid harm are alone sufficient for the conclusion. Instead, we find either the complete silence that is characteristic of the vast majority of ratings publications or, at best, a "just use ratings with caution" statement.

Learning, Customer Satisfaction, and Academic Freedom

Suppose we grant that there is a prima facie institutional moral duty to evaluate faculty teaching in making faculty employment decisions. And, if teaching "effectiveness"—whatever that is—is relevant to employment, then it would follow that improving teaching effectiveness should be strongly correlated with improvement in student learning. (This means demonstrable learning, of course, not student opinions about their learning.) But, despite decades of use of student ratings in employment decisions, it has been an open secret in the research literature that student ratings are not correlated with student learning (Armstrong, 1998; Boring et al., 2016; Olivares, 2003; Uttl et al., 2017). This unhappy fact immediately poses the question, "How can we morally defend the use of student ratings in employment decisions when the alleged measure of teaching effectiveness offers no promise of increased student learning?"

Moreover, even if there were a connection between student learning and teaching effectiveness—as measured by student ratings—that relationship would still be problematic. As Michael Scriven observes, "The best teaching is not that which produces the most learning, since what is learned may be worthless" (Scriven, 1981, p. 248). Put another way, the effectiveness of teaching is not the same thing as the quality of teaching. This distinction deserves serious consideration beyond this paper.

Let's look at a counterargument available to administrators and governing boards regarding the lack of a correlation between ratings and student learning. The institution might choose to place the highest value on student *satisfaction* with a given faculty member and assert that satisfaction is something that student ratings can measure. Two serious problems immediately arise with this counterargument. First, prioritizing student satisfaction over student learning is in clear conflict with one of the primary missions of the university, viz., student learning. Second, in treating student ratings as customer satisfaction surveys, the institution is violating the core ethical principle of public benefit.[13]

The much-discussed consumerist or "student as customer" view of higher education is not new, dating well before World War II. Remmers (1929, p. 7) is perhaps the first ratings researcher to explicitly refer to students as "consumers" of education, and his interest in ratings is guided by that perspective rather than measuring teaching competence:

[12] See Faden and Beauchamp (1986, p. 339) for a widely adopted definition of coercion in a research context.

[13] A thorough discussion of the conflict between the "consumerist" view of higher education and the traditional institutional commitment to public benefit is beyond the scope of this paper.

> I do not pretend to know at present whether or not the pooled judgments of a class do in fact correspond to what might be the objective facts—if such were obtainable—concerning an instructor's competence. Nor am I primarily concerned about this point. The important fact is that student attitudes toward the instructor are certainly of considerable importance in the learning-teaching relationship. It is these attitudes which the scale [*Purdue Rating Scale for Instructors*] is designed to measure, and *not what the instructor is in philosophical actuality* [emphasis added]. (Remmers, 1929, p. 16)

While Remmers can hardly be criticized for his frankness here, such an early and overt abandonment of student ratings as a measure of teaching competence deserves our attention.

Here is another early treatment of students as consumers, leading to the conclusion that students do not need to know much about good teaching in order to respond to a ratings questionnaire:

> The students are the consumers of teaching, and they know what they can and cannot consume, even if they are foggy about the reasons. Students admittedly cannot analyze teaching ability into its elements nor do they often have a clear standard of what constitutes good teaching, but they do not need to have either. They can answer specific questions about their own reactions, and that is all any scale asks them to do. (Cole, 1940, p. 572)

Both Remmers and Cole clearly regard student ratings as opinion polls, although like his contemporary, Guthrie, Remmers conflates the objective evaluation of teaching with subjective student opinion. As Cole implies, one obviously cannot move logically from a collection of student beliefs or opinions to any external facts whatever, other than statistical statements about those very beliefs, thus—again—mooting any alleged value of ratings in faculty employment decisions.

Given the student-as-consumer view of ratings, i.e., where ratings are simply polls of student satisfaction with instruction, ratings are conclusively *uncoupled* from the evaluation of faculty. This uncoupling provides an argument for ending the use of student satisfaction ratings because the data from the instruments in question do not reflect reasoned, evaluative judgments about teaching. Consequently, they must never be used administratively to judge teaching quality.

The ethical consequences of the consumerist approach to undergraduate education reflected by the administration of student satisfaction polls are quite significant. The most prominent defect of the consumerist approach to higher education is that "[c]ustomers want to have their preferences *satisfied*, but students come to a university to have their preferences *formed*" (Kirp, 2003, p. 123). This means that student ratings qua satisfaction polls will intrude into pedagogical choices, thus compromising academic freedom, professional autonomy, and clinical independence.

Jordan Titus' paper, "Student Ratings in a Consumerist Academy" (2008) found that student ratings are

> normative assessments of a professor's conformity with students' pedagogical role expectations that have been derived from a market ideology and framed by a transmission model of education embedded in the rating form. (p. 397)

This conclusion up-ends the common assumption in ratings research that student ratings are somehow objective measurements of external artifacts or "data." Rather, Titus finds that students are bringing consumerist expectations to the classroom. If the transaction meets their expectations, then it will be rated as an enjoyable experience. Second, Titus found that the instrument under study—the University of Washington's *Instructional Assessment System*—was predicated on the typical theoretical SET approach, viz., the transmission ("filling a pail") model of education. Consequently, critical or transformative pedagogies found in various disciplines—sociology in Titus' case—do not rate as highly as do traditional lecture-based pedagogies.

Both of these findings are ethically troubling. First, faculty and administrators who promote the use of student ratings, qua satisfaction polls, in employment decisions are implicitly signing on to a customer-satisfaction model of excellence in teaching. "Truth in advertising" is consequently a necessary condition for moral permissibility here. That is, the institution's commitment to the consumer model must be made explicit in faculty job postings, faculty handbook language regarding retention, tenure, promotion, contract renewal, and merit pay, and on the ratings instrument itself. Plain faculty handbook language is needed. For example, a faculty handbook might state: "Faculty employment decisions will be made wholly, or in large part, on how well the faculty member meets student expectations and on the level of student satisfaction with the faculty member's courses." This is not a facetious proposal. It is a bare paraphrase of a recent recommendation by Uttl et al. (2017):

> Universities and colleges focused on student learning may need to give minimal or no weight to SET ratings. In contrast, universities and colleges focused on students' perceptions or satisfaction rather than learning may want to evaluate their faculty's teaching using primarily or exclusively SET ratings, emphasize to their faculty members the need to obtain as high SET ratings as possible (i.e., preferably the perfect ratings), and systematically terminate those faculty members who do not meet the standards. (p. 40)

For faculty employment processes to be morally acceptable, the conditions of employment must be clearly stated. Of course, this may make recruitment and retention of excellent faculty more difficult for the consumerist university, but that is a separate matter.

Second, the institutional premise of the typical SET qua satisfaction poll seems to be something like the following:

> "Teach your students however you want, but you will be judged by them on how well you comport with pedagogies that do not press them, do not make them uncomfortable in their opinions, and do not challenge them to change their perspectives that they may have brought with them to the course. Comfortable satisfaction is the standard, not any engagement with the instructor that is troubling, unsettling, or, of special concern, challenging. When your students' satisfaction with you has been tabulated, your colleagues will make a judgment about your future at this institution."[14]

Titus (2008) draws the conflict here as one between the "comfortable satisfaction" of the student and challenging students to think critically (403). He describes this as a "distortion"

[14] I am indebted to Slevin (2002, p. 70) for the apt descriptors, "troubling," "unsettling," and "challenging."

of the teacher-student relationship into a market transaction (399). It is difficult to imagine a deeper or more subversive attack on academic freedom.[15]

Given the view that the student is a customer, it is no surprise that consumerist universities would conduct polls of student satisfaction regarding the "product" that students have purchased. This allows such institutions to market their educational products in ways that are most likely to attract new customers. And, well-designed student ratings can certainly provide evidence of such customer satisfaction as well as satisfying what Reis and Klotz describe as the "audit culture" of student ratings in their 2011 paper, "The Road to Loss of Academic Integrity is Littered with SET: A Hypothetical Dilemma."

Even if we were to grant the consumerist model of higher education, it doesn't follow that the quantitative measurement of student satisfaction with their professors constitutes an evaluation of teaching quality. This is because student opinion polls are not measurements, direct or indirect, of anything other than student beliefs. Schueler (1988) puts the point this way:

> [teacher] evaluation polls are "valid" and in fact can be shown to be. There is no reason to doubt that this is so, as long as we remember that "valid" here just means "accurately reflect student opinion" and nothing more. In particular it cannot mean "correctly evaluates this professor as a teacher." That is not something that an opinion poll could be "shown" to do…[N]o poll can tell whether the beliefs it records are correct. To think it could do that would be like thinking that we could discover the nutritional value of some food, say, by conducting a poll of grocery store customers. (p. 346)

The conclusion to be drawn here is that there is no moral justification of using student opinion polls in faculty employment decisions. How could there be? The evaluation of faculty teaching is logically unrelated to student expressions of satisfaction. Every university that currently uses student ratings in faculty employment decisions thus has to address the question, "What kind of institution do we want to be?" Most faculty would prefer to work at a university where the answer is, "Student learning is the most important objective in the classroom." But suppose the answer is, "Yes, we understand that student ratings are not connected with student learning, but learning is only a secondary objective on this campus. We will evaluate faculty on the basis of customer satisfaction because we need to maximize customer enrollment."

In this case, faculty—especially nontenured faculty and faculty applying for promotion or merit pay—must choose between a professional Scylla and Charybdis. They can use their best professional judgment to optimize their students' learning, thereby perhaps making students uncomfortable and so endangering their ratings, or they can select empirically grounded techniques that will maximize student satisfaction. Faculty who choose the latter horn of the dilemma are well-advised to follow the guidelines in Ian Neath's (1996) evidence-based paper, "How to Improve Your Teaching Evaluations without Improving Your Teaching."

[15] Haskell (1997a, b, c, d) provides an extended treatment of how student ratings interfere with academic freedom. Haskell's review of case law regarding the distinction between academic freedom of speech and academic freedom of pedagogical choices is especially useful (see 1997d).

The belief expressed by state legislatures and university governing boards that the university "should be run like a business" reinforces the idea that university teaching is not a true profession. As this meme has embedded itself in higher education, it has brought in its trail various practices that smack of the marketplace rather than the academy. It follows that the patois of corporate management, "branding," "accountability," academic "products," and of course, "customers," cannot simply be dismissed as an inconsequential historical artifact with no critical impact on university teaching. Rather, it undermines the foundations of the profession itself.

In contrast, the concept of teaching as scholarship is a core component of the traditional professions. Morehead and Shedd argue that "teaching should be evaluated, like research, through a peer review process." Citing the AAHE's national study on peer review of teaching (Hutchings et al., 1995), they state that valuing teaching as scholarship allows faculty "to act as a community of scholars" (Morehead & Shedd, 1997). Morehead and Shedd proceed to make a case for external peer evaluation, but the larger point is critical, viz., that the scholarship of teaching *demands* peer review of teaching. The reason for this is that teaching is a domain of expertise not accessible to laypersons outside of university teaching, and certainly not accessible to undergraduate students any more than expertise in the practice of medicine is accessible to and evaluable by medical laypersons.

Conclusion

In summary, the use of student ratings in faculty employment decisions fails to meet ethical scrutiny in three basic ways. First, there is not even an informal consensus definition of teaching effectiveness, the very attribute ratings are alleged to measure or evaluate. We don't have a consensus view about what student ratings are, e.g., whether they are polls of student satisfaction, or instead, are observational or experimental studies of judgments made by competent observers or evaluators of a faculty member's clinical skills. If student ratings are records of evaluative judgments, the matter of student-evaluator competence arises, the only extant defense of which is the fallacious "long exposure" argument. Second, wrongful employment decisions made on the basis of ratings, even in part, are serious harms. University endorsement or encouragement of student participation in an activity that may cause harm to faculty is not ethically defensible. Third, if ratings are merely polls of student satisfaction, then no conclusions about the professor's teaching expertise can be drawn. An evaluation of university teaching grounded in student satisfaction also violates the clinical independence and academic freedom of the instructor, as well as professional ethics standards at both the individual and institutional levels. These ethical failures are each independently sufficient for eliminating the administrative use of student ratings in faculty employment decisions.

Acknowledgements I would like to thank the editor and the anonymous reviewers for their helpful comments. I also appreciate the comments of Louis Clark, Virginia Gregg, Marc Moreau, and Glenn Pascal on an earlier version of this paper.

Author Contributions N/A.

Funding N/A.

Data Availability N/A.

Code Availability N/A.

Declarations

Conflicts of Interest No conflicts of interest.

Ethics Approval N/A.

Consent to Participate N/A.

Consent for Publication N/A.

References

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology*, *82*(2), 219–231. https://doi.org/10.1037/0022-0663.82.2.219

Aleamoni, L. M. (1987). Typical faculty concerns about student evaluation of teaching. *New Directions for Teaching and Learning*, *1987*(31), 25–31. https://doi.org/10.1002/tl.37219873105

Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*(2), 153–166. https://doi.org/10.1023/A:1008168421283

American Education Research Association, American Psychological Association, & National Council on Measurement in. (2014). In Education (Ed.), *Standards for educational and psychological testing*. American Education Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/9780935302356.pdf

Anthology (2022). IDEA's history. https://www.ideaedu.org/about-idea/ideas-history/

Armstrong, J. S. (1998). Are student ratings of instruction useful? *American Psychologist*, *53*(11), 1223–1224. https://repository.upenn.edu/handle/20.500.14332/39404

Beauchamp, T. L., Walters, L., Kahn, J. P., & Mastroianni, A. C. (2008). *Contemporary issues in bioethics* (7th ed.). Wadsworth.

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, *January 7*. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Brandenburg, G. C., & Remmers, H. H. (1927). The Purdue Rating Scale for instructors. *Educational Administration and Supervision*, *13*, 399–406.

Brandenburg, G. C., & Remmers, H. H. (1928). *Manual for the Purdue Rating Scale for Instructors*. Lafayette Printing.

Campus Labs (2020). Getting started - IDEA instruments. https://courseevaluationsupport.campuslabs.com/hc/en-us/articles/360038358293-Getting-Started-IDEA-Instruments

Close, D. (1976). What is non-epistemic seeing? *Mind*, *85*(338), 161–170. https://doi.org/10.1093/mind/LXXXV.338.161

Cohen, P. A. (1990). Bringing research into practice. *New Directions for Teaching and Learning*, *1990*(43), 123–132. https://doi.org/10.1002/tl.37219904311

Cole, L. (1940). *The background for college teaching*. Farrar & Rinehart.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist*, *52*(11), 1198–1208. https://doi.org/10.1037/0003-066X.52.11.1198

U. S. Department of Health and Human Services (2020). *Human subject regulations decision charts: 2018 requirements*. https://www.hhs.gov/ohrp/regulations-and-policy/decision-charts-2018/index.html

Detchen, L. (1940). Shall the student rate the professor? *The Journal of Higher Education*, *11*(3), 146–154. https://www.jstor.org/stable/1974002

Donovan, A. (2013). Why should evaluators care about research ethics? Solutions IRB. https://solutionsirb.com/wp-content/uploads/2013/05/Why-should-evaluators-care-IRB.pdf

Doyle, K. O. (1983). *Evaluating teaching*. Lexington Books.

Dretske, F. L. (1969). *Seeing and knowing*. University of Chicago Press.

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

Faden, R. R., & Beauchamp, T. L. (1986). *A history and theory of informed consent*. Oxford University Press.

Flaherty, C. (2018). Teaching eval shake-up. *InsideHigherEd*, May 21. https://www.insidehighered.com/news/2018/05/22/most-institutions-say-they-value-teaching-how-they-assess-it-tells-different-story

Franklin, J., & Theall, M. (1989). Who reads ratings: Knowledge, attitude, and practice of users of student ratings of instruction. Paper presented at the 70th annual meeting of the American Educational Research Association, San Francisco, CA: March 31. https://files.eric.ed.gov/fulltext/ED306241.pdf

Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models, and trends. Toronto: Higher Education Quality Council of Ontario*. Student Course Evaluations.pdf. https://www.heqco.ca/SiteCollectionDocuments/

Guthrie, E. R. (1927). Measuring student opinion of teachers. *School and Society*, *25*(February 5), 175–176.

Guthrie, E. R. (1954). *The evaluation of teaching: A progress report*. University of Washington.

Hashimoto, N. (2006). Professional autonomy. *Japan Medical Association Journal*, *49*(3), 125–127. https://www.med.or.jp/english/pdf/2006_03/125_127.pdf

Haskell, R. E. (1997a). Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century. *Education Policy Analysis Archives*, *5*(6), 1–31. https://doi.org/10.14507/epaa.v5n6.1997

Haskell, R. E. (1997b). Academic freedom, promotion, reappointment, tenure, and the administrative use of student evaluation of faculty: (part II) views from the court. *Education Policy Analysis Archives*, *5*(17), 1–44. https://doi.org/10.14507/epaa.v5n17.1997

Haskell, R. E. (1997c). Academic freedom, promotion, reappointment, tenure and the administrative use of student evaluation of faculty (SEF): (part III) analysis and implications of views from the court in relation to accuracy and psychometric validity. *Education Policy Analysis Archives*, *5*(18), 1–44. https://doi.org/10.14507/epaa.v5n18.1997

Haskell, R. E. (1997d). Academic freedom, promotion, reappointment, tenure, and the administrative use of student evaluation of faculty (SEF): (part IV) analysis and implications of views from the court in relation to academic freedom, standards, and quality instruction. *Education Policy Analysis Archives*, *5*(21), 1–42. https://doi.org/10.14507/epaa.v5n21.1997

Hutchings, P., Teaching Initiative, A. A. H. E., & Stanford University. (1995). *From idea to prototype: The peer review of teaching: A project workbook*. Stylus Publishing.

ICES (University of Illinois Urbana-Champaign Center for Innovation in Teaching & Learning) (2023). Teaching evaluation: ICES myths & misperceptions. https://citl.illinois.edu/citl-101/measurement-evaluation/teaching-evaluation/teaching-evaluations-(ices)/myths-misperceptions

Kampmeier, R. H. (1972). The Tuskegee study of untreated syphilis. *Southern Medical Journal*, *65*(10), 1247–1251. https://doi.org/10.1097/00007611-197210000-00016

Kirp, D. L. (2003). *Shakespeare, Einstein, and the bottom line: The marketing of higher education*. Harvard University Press.

Kreitzer, R. J., & Sweet–Cushman, J. (2022). Evaluating student evaluations of teaching: A review of measurement and equity bias in SETs and recommendations for ethical reform. *Journal of Academic Ethics*, *20*(March), 73–84. https://doi.org/10.1007/s10805-021-09400-w

Kulik, J. A., & McKeachie, W. J. (1975). The evaluation of teachers in higher education. *Review of Research in Education*, *3*, 210–240. https://doi.org/10.2307/1167259

Li, D., Benton, S. L., Brown, R., Sullivan, P., & Ryalls, K. R. (2016). *IDEA technical report no. 19: Analysis of IDEA student ratings of instruction system 2015 pilot data*. Manhattan, KS: IDEA. https://www.ideaedu.org/Portals/0/Uploads/Documents/Technical-Reports/IDEA_Technical_Report_No_19.pdf

Lindahl, M. W., & Unger, M. L. (2010). Cruelty in student teaching evaluations. *College Teaching*, *58*(3), 71–76. https://doi.org/10.1080/87567550903253643

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

Marsh, H. W. (1982). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology*, *52*(1), 77–95. https://doi.org/10.1111/j.2044-8279.1982.tb02505.x

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

Marshall, M., & Clark, A. M. (2010). Is clarity essential to good teaching? *Teaching Philosophy*, *33*(3), 271–289. https://doi.org/10.5840/teachphil201033329

Merritt, D. J. (2008). Bias, the brain, and student evaluations of teaching. *St John's Law Review*, *82*(1), 235–287. https://scholarship.law.stjohns.edu/cgi/viewcontent.cgi?article=1100&context=lawreview

Michigan State University Board of Trustees. (2011). *SIRS online student instructional rating system: Frequently asked questions*. https://sirsonline.msu.edu/FAQ.asp#id-2

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*(4), 371–378. https://doi.org/10.1037/h0040525

Milgram, S. (1973). The perils of obedience. *Harper's Magazine*, *247*(1483), 62–78.

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

Morehead, J. W., & Shedd, P. J. (1997). Utilizing summative evaluation through external peer review of teaching. *Innovative Higher Education*, *22*(1), 38. https://doi.org/10.1023/A:1025199425293

National Council of Measurement in Education. (2016). Code of professional responsibilities in educational measurement. *Assessment in Education: Principles, Policy & Practice*, *3*(3), 401–412. https://doi.org/10.1080/0969594960030309

National Research Act. (1974). 42 U.S.C. § 289.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. U.S. Department of Health and Human Services. https://videocast.nih.gov/pdf/ohrp_belmont_report.pdf

Neath, I. (1996). How to improve your teaching evaluations without improving your teaching. *Psychological Reports*, *78*(3_suppl), 1363–1372. https://doi.org/10.2466/pr0.1996.78.3c.13

Office of Human Research Protections. (2017). Federal policy for the protection of human subjects. *Federal Register*, *82*(12), 7149–7274. https://www.federalregister.gov/documents/2017/01/19/2017-01058/federal-policy-for-the-protection-of-human-subjects

Olivares, O. J. (2003). A conceptual and analytic critique of student ratings of teachers in the USA with implications for teacher effectiveness and student learning. *Teaching in Higher Education*, *8*(2), 233–245. https://doi.org/10.1080/1356251032000052465

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Institutional Research*, *2001*(109), 27–44. https://doi.org/10.1002/ir.2

Protection of Human Subjects. (2009). 45 C.F.R. § 46. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/#46

Reis, J., & Klotz, J. (2011). The road to loss of academic integrity is littered with SET: A hypothetical dilemma. In *Educational integrity: Culture and values. Proceedings 5th Asia Pacific Conference on Educational Integrity*. The University of Western Australia, September 26–28 (pp. 110–120). https://www.academia.edu/974502/The_road_to_loss_of_academic_integrity_is_littered_with_SET_A_hypothetical_dilemma

Remmers, H. H. (1929). The college professor as the student sees him. *Bulletin of Purdue University: Studies in Higher Education XI*, *29*(6), 1–63.

Remmers, H. H. (1933). Learning, effort, and attitudes as affected by three methods of instruction in elementary psychology. *Bulletin of Purdue University: Studies in Higher Education XXI*, *33*(6), 1–48.

Remmers, H. H. (1958). On students' perceptions of teachers' effectiveness. In W. J. McKeachie (Ed.), *The appraisal of teaching in large universities: A report* (pp. 17–23). The University of Michigan.

Rowan, S., Newness, E. J., Tetradis, S., Prasad, J. L., Ko, C. C., & Sanchez, A. (2017). Should student evaluation of teaching play a significant role in the formal assessment of dental faculty? Two viewpoints: Viewpoint 1: Formal faculty assessment should include student evaluation of teaching and viewpoint 2: Student evaluation of teaching should not be part of formal faculty assessment. *Journal of Dental Education*, *81*(11), 1362–1372. https://doi.org/10.21815/JDE.017.093

Schueler, G. F. (1988). The evaluation of teaching in philosophy. *Teaching Philosophy*, *11*(4), 345–348. https://doi.org/10.5840/teachphil198811485

Scriven, M. (1981). Summative teacher evaluations. In J. Milliman (Ed.), *Handbook of teacher evaluation* (pp. 244–271). Sage.

Scriven, M. (1995). Student ratings offer useful input to teacher evaluations. *Practical Assessment Research & Evaluation*, *4*(7). https://doi.org/10.7275/1jfr-et33

Shils, E. (1983). *The academic ethic*. Chicago, IL: The University of Chicago Press. Reprinted from The academic ethic: The report of a study group of the International Council on the Future of the University. *Minerva*, *20*(1/2), 1982, 107–208. https://www.jstor.org/stable/41820489

Shook, B., & Greer, M. (2015). The misanalysis, misinterpretation, and misuse of student end-of-course evaluation data. *National Social Science Journal*, *44*(2), 89–97.

Slevin, J. F. (2002). Keeping the university occupied and out of trouble. *ADE Bulletin*, *130*(Winter), 63–71. https://www.jstor.org/stable/25595731

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

Stark, P. B. (2016). *Expert report on student evaluations of teaching (Faculty course surveys)*. Prepared for the Ryerson Faculty Association and the Ontario Confederation of University Faculty Associations. October 10. https://ocufa.on.ca/assets/RFA.v.Ryerson_Stark.Expert.Report.2016.pdf

Sullivan, G. M. (2011). Education research and human subject protection: Crossing the IRB quagmire. *Journal of Graduate Medical Education*, *3*(1), 1–4. https://doi.org/10.4300/JGME-D-11-00004.1

Theall, M. (1990). J. Franklin (Ed.), Editors' notes. *New Directions for Teaching and Learning* 1990 43 1–4 https://doi.org/10.1002/tl.37219904302

Theall, M. (2003). Student ratings: Myths vs. research evidence. *Teaching Forum*, Fall. https://www.vanderbilt.edu/cft/resources/cft_newsletters/fall2003/student_ratings_theall.htm

Theall, M., Abrami, P. C., & Mets, L. A. (Eds.). (2001). The student ratings debate: Are they valid? How can we best use them? *New Directions for Institutional Research, 2001*(109).

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research*, *2001*(109), 45–56. https://doi.org/10.1002/ir.3

Titus, J. J. (2008). Student ratings in a consumerist academy: Leveraging pedagogical control and authority. *Sociological Perspectives*, *51*(2), 397–422. https://doi.org/10.1525/sop.2008.51.2.397

University of Chicago. (1926). List of qualities desirable in instructors in elementary courses conducted by the lecture-discussion method. *Report of the Better Yet faculty-student committee on the quality of instruction in elementary courses*. In Local and chapter notes, Jacob Viner, *Bulletin of the American Association of University Professors (1915–1955)*, *12*(4), 235–241. https://www.jstor.org/stable/40217481

University of Illinois at Urbana-Champaign. (2022). *Instructor and course evaluation system (ICES)*. https://citl.illinois.edu/citl-101/measurement-evaluation/teaching-evaluation/teaching-evaluations-(ices).

Uttl, B., & Smibert, D. (2017). Student evaluations of teaching: Teaching quantitative courses can be hazardous to one's career. *PeerJ*, *5*, e3299. https://doi.org/10.7717/peerj.3299

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*(September), 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Vasey, C., & Carroll, L. (2016). How do we evaluate teaching? Findings from a survey of faculty members. *Academe*, *102*(3). https://www.aaup.org/article/how-do-we-evaluate-teaching

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment & Evaluation in Higher Education*, *29*(2), 191–212. https://doi.org/10.1080/0260293980230207

Warnock, G. J. (1965). Seeing. Reprinted with Postscript, 1963. In R. J. Swartz (Ed.), *Perceiving, sensing and knowing* (pp. 49–67). Anchor Books.

Wilson, W. R. (1932). Students rating teachers. *The Journal of Higher Education*, *3*(2), 75–82. https://www.jstor.org/stable/1975183

Wotruba, T. R., & Wright, P. L. (1975). How to develop a teacher-rating instrument: A research approach. *The Journal of Higher Education*, *46*(6), 653–663. https://doi.org/10.2307/1979060