Why it is wrong to use student evaluations of professors as a measure of teaching effectiveness in personnel assessments: an unjust risk of harm account

Aloyo, E.T.

Public Affairs Quarterly
Volume 37, Number 2, April 2023

WHY IT IS WRONG TO USE STUDENT EVALUATIONS OF PROFESSORS AS A MEASURE OF TEACHING EFFECTIVENESS IN PERSONNEL ASSESSMENTS: AN UNJUST RISK OF HARM ACCOUNT

Eamon Aloyo

I argue that university supervisors should not use student evaluations of teachers (SETs) as a measure of teaching effectiveness in personnel assessments because the evidence suggests SETs likely violate several duties university supervisors have toward their instructional employees. I focus on the duty to not knowingly impose a wrongful risk of harm on nonconsenting and innocent others. Many university employers impose a wrongful risk of harm on instructors by not using relevant, merit-based performance indicators that have adequate construct validity, by using uncorrected indicators that likely perpetuate discrimination, and by incentivizing instructors to do wrong. The use of SETs imposes unjust risk of harm on all instructors, but the risk is higher for women, minorities, and those in precarious, teaching-focused roles. In conclusion, I tentatively suggest some other means of evaluating student learning and assessing instructors.

Keywords: Student Evaluations of Teachers (SETs), risk, harm, duty, discrimination

Introduction

At least sixteen thousand universities and colleges across the globe now commonly use self-reported student evaluations of teachers (SETs).[1] Students typically fill out SETs anonymously at the end of a course. Instructors usually receive summaries of numeric indicators and open-ended written feedback. Student evaluations of

teachers generally include questions on the quality of the instructor, the quality of the course, how much students say they learned, and other questions. What is not included in SETs is important: they typically contain no objective measures of how much students learned in a course either on the specific topic of a course or how much they improved a more general skill (e.g., critical thinking). Student evaluations of teachers have at least two uses. One aim of SETs is to help teachers improve their courses. A second common use of SETs is for supervisors to assess instructors. Furthermore, these assessments are widely used in personnel evaluations in combination with other factors, especially research productivity. Is the second use of SETs justifiable? My central claim is that the use of SETs to assess teaching effectiveness is morally impermissible in personnel assessments because doing so violates several duties supervisors have toward instructors, especially the duty not to impose an unjust risk of harm on instructors.

This paper's argument unfolds as follows. I argue that if an employer fails to discharge several key duties, then she exposes an employee to a wrongful risk of harm. As a specific case of a general duty, employers should not expose (non-consenting and innocent) employees to a wrongful risk of harm. Specifically, building on the work of Jason Brennan and Phillip Magness,[2] I argue that employers have several duties relating to treating their employees fairly. These include using only relevant, merit-based performance metrics that have reasonable construct validity, that is, a reasonable likelihood of adequately measuring a concept an indicator is supposed to represent, as Brennan and Magness argue.[3] Supervisors have a duty to not incentivize employees to violate any of their duties. Supervisors have another duty to only use performance metrics that have a reasonable likelihood of not being biased against individuals who are members of disadvantaged groups, or adjusting the indicators for individuals who are members of disadvantaged groups so that they are likely to be roughly equivalent to others' scores. I then again present evidence demonstrating that using SETs in performance evaluations of instructor effectiveness violates each duty mentioned above. As a result, university administrators impose an unjust risk of harm on instructors. Next, I suggest some tentative solutions for how supervisors can use other indicators to assess teacher effectiveness. A caveat is that some student feedback may be useful for ends other than assessing teaching effectiveness, such as for feedback to instructors, but this should be balanced against the harm caused by sexist and other types of inappropriate comments that sometimes appear in the open-ended component of SETs.

Duties Related to University Instruction and a Duty Not to Impose a Wrongful Risk of Harm

What are the duties university supervisors owe their instructors in assessing their teaching? What duties do professors owe their students regarding

instruction? And how do these duties interact and relate to SETs? Brennan and Magness propose two criteria that university supervisors should use to evaluate instructors.[4] One is a supervisor's "obligation to faculty to evaluate them on the basis of fair and reasonable criteria."[5] The other is a supervisor's "obligation to students to use fair and reasonable criteria in determining who their teachers will be."[6] Brennan and Magness argue that SETs violate both of these principles and therefore university supervisors are wrong to use them to evaluate teaching staff.[7] Their views are plausible and persuasive. However, their arguments are incomplete.

My account of why it is wrong to use SETs in performance evaluations of teachers differs in several important ways from Brennan and Magness's argument. First, I provide a more complete catalog than Brennan and Magness of the duties that should be operative in universities, including which duties university supervisors owe their employees, which duties instructors owe their students, and how these duties interact. This more complete assessment of duties provides grounds for a novel argument why using SETs as a performance metric of teaching effectiveness is wrong: an excess and unnecessary risk of harm to instructors. This harm is especially wrong because it is unevenly distributed, imposing greater excess risk on already disadvantaged individuals.

I build this argument by showing how violating other duties imposes a wrongful risk of harm. One duty Brennan and Magness omit is a detailed discussion of a supervisor's duty to not perpetuate discrimination. This duty is important in itself. The duty to not perpetuate discrimination is also instrumentally important because of the implications for the construct validity of SETs and its implications for the probability of using SETs in personnel assessments harming instructors. If there is evidence that members of socially salient groups who are typically discriminated against in society, such as women, black and brown people, and others, receive lower SETs, but there is no evidence that members of such groups are worse teachers, this is another piece of evidence that SETs do not have adequate construct validity. I consider a duty that instructors have toward their students, namely, to use the teaching methods that are likely to help students learn the most, if it is not too costly to do so. Supervisors have a duty to not incentivize wrongdoing of their employees, at least if it is not too costly to do so. These two duties interact in relation to SETs because SETs likely incentivize some instructors to use teaching methods that the evidence suggests do not help students learn the most. The account I offer here demonstrates why violating these duties is wrong in terms of imposing an excess and unnecessary risk of harming employees. I begin with this discussion of risk.

It is widely accepted that everyone has a duty to not impose a wrongful risk of harm on others. This duty holds even if in any instance, no one is actually harmed. Common examples include the duty to not drive drunk or the duty to refrain from actions such as firing a gun for fun in an urban setting, both of which could kill

or seriously injure an innocent person and can easily be avoided. In this section, I develop the argument that (university) employers also have an employment-related duty to not impose a wrongful risk of harm on their employees, especially if this can easily be avoided, but that using SETs as measures of teaching effectiveness violates this and other duties.

How to define harm has generated a rich literature, and it is beyond the scope of this paper to contribute to those debates. More than one definition of harm is compatible with the claims in this paper. The account of harm I adopt here is a Millian one in which people's important interests are not adequately respected.[8] The interests in question include a panoply of employment-related goods including income and associated benefits such as health care insurance coverage (in the United States), retirement benefits, self-respect, emotional and psychological well-being, and freedom from situations in which one must select some wrongdoing or else risk worse performance evaluations.

The employer's and supervisor's role-based duties to not impose a wrongful risk of harm are derived from a general duty for everyone to not impose a wrongful risk of harm on others, at least if it can be easily avoided. The reason I focus on the *risk* of harm is that this captures cases in which someone may not actually be harmed, or we cannot know whether someone is actually harmed, but where the risk of harm is still wrong because it is excessive and based on irrelevant factors. Risk is the chance of some morally weighted negative event occurring.[9] What makes a certain type of risk of harm wrongful? There could be multiple ways that a risk of harm becomes wrongful, such as imposing excess or unnecessary severe risk on innocent, non-consenting others. Risk can be excessive when holding the type of harm constant, the severity of harm could increase when holding the risk of it constant, or both the risk and severity of harm could increase. These ideas are often captured through concepts such as recklessness and negligence. A university supervisor could violate her duty to not impose a wrongful risk of harm in at least several ways. I motivate and discuss these duties in turn: a duty to only use relevant and valid performance indicators, a duty not to likely perpetuate discrimination, and a duty to not incentivize wrongdoing.

A Duty to Only Use Valid and Relevant Performance Indicators

A performance evaluation measure should not be arbitrary. As Robert Audi states, "any reasonable employment policy . . . depends on conceptions of qualifications and merit."[10] I start with the assumption that employers treat their employees fairly only when they use a relevant, merit-based indicator in employee assessments and personnel decisions. To be relevant, a merit-based performance indicator should closely track the duties of a job.[11] Performance indicators should not be arbitrary

or result in arbitrary assessments. When assessing professors on their teaching, arguably, the most important merit is instructional ability. This, in turn, should be measured primarily by how much students learn. There could be additional teaching related merit indicators. These could include how much a professor inspires students to pursue worthy goals, pushes students to do their best work, or influences students to become lifelong learners. Such goals illustrate the difficulty of measuring relevant indicators and developing alternatives to SETs. But whatever other instructional merit indicators one considers valuable, it is only plausible that student learning must be a centerpiece of any educational institution and hence, of any instructor's work.

Other factors besides an instructor's teaching expertise nearly certainly influence how much students learn.[12] Student-based factors that might influence how much a student learns in a course include a student's motivation, intelligence, and prior knowledge of the subject. Institutional factors might influence how much students learn too. These could include whether a professor teaches a mandatory introductory class or an upper-level seminar, or whether the instructor has control over a syllabus. Only by controlling for such factors would one isolate the independent influence a professor has on student learning. Therefore, measuring student learning without controlling for other factors should not be the sole way of assessing an instructor's competence.

Another aspect of the first duty is that supervisors should use indicators that adequately measure the concepts on which instructors should be assessed. What do I mean by adequately measure a concept? By this, I mean that the best available evidence suggests any indicator in question is correlated with student learning. It is a fatal flaw of a measure if the evidence is so noisy that there is no consistent correlation between it and student learning, or if there is an inverse correlation between student learning and the performance indicator. Any indicator used to measure teaching ability should at least roughly approximate this concept. This is called validity,[13] or, more precisely, construct validity.[14] The higher the construct validity, the more closely an indicator measures the concept in question. Student evaluations of teachers are supposed to have high construct validity for instructor effectiveness. Student evaluations of teachers use questions such as "How much did you learn in this course?" ostensibly to assess how much students learn. This has high face validity. But I will present evidence to show why the claim that SETs measure learning is almost certainly false.

One might make the objection that invalid measures are fair if everyone were subject to the same risk. But subjecting employees equally to some risk that is unrelated to what a fair means of assessing an employer is wrong if there is a duty to only use fair measures of assessment. Imagine that a sadistic employer enjoyed firing employees if when they entered their annual review, a roll of three dice all landed on sixes. All employees would be subject to the same risk, but in every case, this evaluation method is arbitrary and unrelated to someone's

job performance. This is an unjust risk of harm from an employer because it is arbitrary.

A Duty to Not Likely Perpetuate Discrimination

As Kasper Lippert-Rasmussen puts it, "discrimination in the sense that interests us here is discrimination against a socially salient group or particular individuals *qua* members of a socially salient group."[15] Discrimination can take many forms. I focus on wrongful employment discrimination. What should count as wrongful discrimination? Lippert-Rasmussen offers a harm-based account: "An instance of discrimination is *pro tanto* bad, when it is, because it makes the discriminatees [*sic*] worse off."[16] This is an appealing view because it captures the widely agreed-upon aspect of discrimination that it decreases someone's well-being. Disrespecting someone, paying someone less than an equally qualified colleague, and overlooking someone for a job because of membership in a socially salient group are all-too-common means of wrongful discrimination. Another plausible account suggests discrimination can be wrong when it impedes equality of opportunity.[17] But not all discrimination is wrong if it is defined as above. A person could be worse-off or may not have the same opportunities because of an age-related prohibition. For instance, voting rights and driving rights typically have minimum age requirements. But age requirements alone do not make the restrictions wrong. Then when is discrimination wrong?

At least one way that discrimination can be wrong is because the feature on which someone is discriminated is arbitrary or irrelevant. One's age is likely a relevant feature in deciding when someone should be able to drive, given the irresponsibility and recklessness to which youth (especially male youth) are prone. Age-based policies need not always be wrongfully discriminatory. But whether one is a member of a socially salient group is always or almost always irrelevant for teaching ability.[18] Thus, using performance indicators that perpetuate discrimination because of membership in a socially salient group is *prima facie* wrongful.

An account of discrimination like Lippert-Rasmussen's has concerning features, however. Specifically, it is problematic if some decision intuitively seems like discrimination but is not captured by the definition because it does not leave someone worse-off. Consider a case of harmless discrimination that Tom Parr presents.[19] The case Parr imagines is one in which a hiring committee is biased against immigrants. Instead of simply rejecting an immigrant's application, the hiring committee at firm A uses their connections to get firm B, that the applicant favors, to hire her. She is not harmed because she gets the job that she prefers (at firm B), and which will leave her better-off than the alternative opportunity. Segall's account might seem to fare better here because trying to avoid offering a job to an immigrant candidate impedes equality of opportunity. But per

stipulation, the person who is ostensibly discriminated against in fact has a better opportunity, so it is hard to see how it prevented equality of opportunity. If anything, the person whom the original group decides to hire might be seen to be discriminated against in this case, as the hiring committee did not praise him to the better firm. Surely this cannot be correct. The committee that pulled the strings discriminated against the candidate because of her membership in a social group.

This applies to the discussion of the ethics of SETs in the following way. Suppose, by a stroke of luck, that all of a female instructor's students are completely unbiased, and this is reflected in her SETs. By stipulation, she's no worse-off and has the same opportunity as men because her students showed no bias toward her in her SETs. Would it still be wrong to use SETs in evaluating her? If so, why would it be wrong, even in this case where, collectively, the students exhibited no biases against the professor? Assume further that her supervisor (let us assume) is not biased against women or any other socially salient group. This sort of case is important to consider because we cannot know from the general evidence that any one individual is discriminated against through SETs, even if she receives lower marks than her male counterparts.

My account of wrongful risk of employment-related harm explains why even if one individual is not discriminated against, it remains wrong to use SETs if they generally perpetuate discrimination against individuals who are members of socially disadvantaged groups. At work, it is wrong to expose someone to a worse risk of discrimination than another person for an arbitrary reason. Why is it wrong? It is wrong because employers owe it to their employees to evaluate all of them equally on their merits, but using performance metrics that perpetuate discrimination violates this principle. In other words, even if we cannot know that for any given individual, if the individual is discriminated against, it would still be wrong to use SETs in performance evaluations if there is good reason to believe that SETs systematically discriminate against one more socially salient group.

The account I offer in this article differs from the above accounts because it focuses on the risk of harm. One type of wrongful discrimination is an increase in the probability of a poor job performance assessment for an arbitrary factor that should be unrelated to job performance, such as one's gender, race, sexual orientation, gender identity, native language, and so on. Let us unpack this definition. The claim is not that something like one's native language can never be an important factor in employment decisions. That a candidate is a native Spanish speaker who is applying to be a Spanish teacher is a reasonable factor to consider when hiring. The claim is that when such demographic factors should be irrelevant for employment assessment, they should not give one an advantage or disadvantage in employment. In the overwhelming majority of cases regarding employment and promotion, someone's gender, race, sexual orientation, gender identity, or other socially salient factor is an irrelevant characteristic.

Finally, intention is not required here for wrongful harm. I do not claim that any university supervisor intentionally uses SETs to discriminate against women, black- and brown-skinned people, or any individuals from any other socially salient group. Nor do I claim that anyone intentionally uses SETs to impose risks of harm on instructors for these or other reasons. It is generally worse if someone uses an indicator intentionally to discriminate against an individual who is a member of a group the supervisor dislikes or disrespects, but it need not be a necessary component of imposing a wrongful risk of harm.

One might question whether any increase in the probability of a poor job performance assessment for arbitrary reasons, rather than a substantial increase, should count as wrongful discrimination. I think it should. That is because although greater increases in the probability of and the effective size of a negative performance indicator is worse, any indicator that systematically disadvantages some group of people for an arbitrary reason is wrong. No one should be made worse-off, or risk being worse-off, because of an arbitrary factor such as one's race, gender, or other socially salient feature. But for the sake of argument, I take no position whether any increase in the probability of a poor job performance assessment for arbitrary reasons is wrong. Instead, I will suggest a substantial increase in risk is a non-trivial increase in risk. A substantial increase in the risk of harm for an arbitrary factor imposes a wrongful risk of harm. Without having to quantify that increase here, I will suggest that effect sizes found in empirical research presented below qualified as a substantial, that is, not trivial, increase in risk. Now, what does the evidence suggest about to what extent SETs perpetuate discrimination?

A Duty to Not Incentivize Wrongdoing

An employer could incentivize employees to violate some of their duties toward others. For instance, imagine the police have duties to (among other things) investigate all crimes, and if they do not have the time or other resources to do so, they should focus on investigating the worst crimes. But instead, imagine a police department pressures its staff to make money off traffic stops to maximize revenue.[20] This could wrongly incentivize officers to violate the above duty. University managers also have the duty to not create incentive structures that likely induce employees to violate any of their duties, especially if it is easy to avoid creating such incentives. This derives from a duty to let others discharge their duties.[21] I will argue that using SETs as indicators in performance evaluations violates this duty employers have toward their employees because it incentivizes teachers to use teaching methods that the evidence suggests are not the best way to help students learn the most.

A plausible duty of professors is to adequately educate ourselves on what teaching techniques are most likely to help students learn the most, and employ those techniques in our teaching, assuming they are not overly costly. There

is good evidence that one of the most effective techniques of teaching is using active rather than passive learning.[22] Active learning techniques include activities such as effortful recall that can be operationalized by quizzes spaced out over weeks and months, in contrast with passive learning such as traditional lectures.

Furthermore, the (over)reliance on SETs as measures of teacher performance may contribute to mediocre student achievement in areas such as critical thinking. As Arum and Roksa document, after 2 years of US university-level education, nearly half of students demonstrate no improvement in critical thinking, complex reasoning, or writing, as measured by the Collegiate Learning Assessment (CLA) exam.[23] They present several pieces of evidence to suggest that the amount students learn in the United States has decreased over the last several decades.[24] Although many factors likely contribute to this, instructors may practice more lenient grading to try to achieve higher scores on SETs.[25]

Is it easy for supervisors to avoid creating perverse teaching incentives? Yes. University administrators and supervisors could simply stop using SETs in performance evaluations. For instance, Michael Quick, a provost at the University of Southern California (USC), reportedly decided to simply stop using SETs in promotion and tenure decisions after reviewing the evidence, even though USC still uses them for other purposes.[26]

I argue that supervisors who use SETs to measure teaching effectiveness violate all three duties, and therefore impose a wrongful risk of harm on employees, especially employees who are members of disadvantaged groups. One might object that university employers do not impose a wrongful risk of harm on instructors because instructors consent to being evaluated in part by SETs. Indeed, consent plays a key role in explaining why employers in many dangerous professions do not impose a wrongful risk of harm on their employees. For instance, soldiers knowingly consent to the employment-related risk of being killed or maimed in war. Yet we typically think that militaries of legitimate states impose no wrongful risk of harm in such cases if soldiers must consent to enlist exactly because they consent knowing the broad outlines of the risks to which they may be exposed, and are free to choose other professions.

There are several responses to this objection. First, no one should have to subject oneself to arbitrary and likely discriminatory personnel assessment policies. Although the risk of harms soldiers are subject to is far more severe than the academic, soldiers, too, deserve a non-arbitrary process of performance review and promotion. In other words, the level of harm one is exposed to is morally distinct from the assessment criteria. Second, an instructor cannot opt out of the assessment method if she wants to stay employed at almost any institution of higher education. This objection would be much stronger if a university offers an easy way for an instructor to opt out of SETs being the way their teaching is assessed. But universities never do this, to my knowledge. One necessary aspect

of consent for it to be valid is a choice where the consequences of not consenting are not excessive. Here, the choice one has is to not accept nearly any faculty position if one does not want to be assessed through SETs.

What then does the evidence suggest about to what extent SETs are constructively valid measures of teaching effectiveness, likely involve wrongful discrimination, or incentivize wrongdoing?

Evidence on Construct (In)Validity and the Effectiveness of Teaching Techniques

The burden of proof of using any performance indicator in employee assessments should rest with the persons or institutions who propose or use such instruments. There should be some positive justification for any measure used in personnel assessments. This should apply to each of two steps: selecting a concept that should be used as a performance indicator and operationalizing that concept. The first step concerns which concepts should be included in assessing an employee. For instance, should how much a student learns be an indicator to assess instructor effectiveness? Should how much more (or less) curious a student becomes after taking a course be included as a concept? Teacher effectiveness is a relatively uncontroversial concept that many people take to matter morally and that can be reasonably taken to matter in assessing the quality of an instructor. How much a student learns in a course is similarly important and widely accepted as an important indicator. The second step concerns how those concepts should be measured. This step is often difficult and controversial.

I will assume that the concept of teaching effectiveness should be correlated with actual student learning over sufficiently large numbers of students. Any definition of teaching effectiveness that would not have student learning as a centerpiece of its aim would have something gravely amiss. Teacher effectiveness is not the only factor that influences student learning. Nor should it be the only worthy goal of teaching. Do teachers matter at all for student learning? If not, there might be no use in trying to come up with relevant, merit-based indicators of teaching. Unsurprisingly, there is at least some evidence that skilled teachers can have a positive influence on student achievement.[27]

There is strong evidence to doubt that SETs have adequate construct validity for using them in personnel evaluations. Trivial factors, such as giving cookies to students and the perceived attractiveness of an instructor, influence student ratings of instructors.[28] Several meta-analyses conclude that SETs are a poor measure of student learning.[29] In fact, the authors of a recent and comprehensive meta-analysis conclude "that the SET/learning correlation is zero. . . . Students do not learn more from professors who receive higher SET ratings."[30]

In fact, when instructors use active learning teaching techniques, the evidence suggests that higher markings on SETs are *inversely* correlated with student learning.[31] If instructors know this, using SETs in performance evaluations thereby

incentivizes teachers to not use the techniques that help students learn the most. This is especially problematic if teachers owe a special consideration to those students who are most likely to fail or drop out, because using active learning techniques helps the most disadvantaged students the most.[32]

One objection could be that SETs are in fact relevant, merit-based indicators because they measure student satisfaction (not student learning).[33] Student satisfaction, one might argue, is an important and relevant concept for universities who, after all, need to pay their instructors and other staff and depend on students for revenue. Student evaluations of teachers are easy to use, have face validity, and provide a mechanism for students to provide feedback to instructors on how to improve a course. Such feedback could improve courses by, for instance, identifying shortcomings of a specific course or means of instruction. There are several responses to this objection. One is to simply accept that SETs could have some legitimate uses such as feedback to instructors. As Albert Hirschman argues,[34] voice is an important means for individuals within institutions to express their feedback. This is especially true for students because Hirschman's alternative, exit, is costly, either considering dropping out of university, a major, or a course (especially if it is a required course and often or always taught by the same professor). Student evaluations of teachers provide a useful means for students to express their views about a course and instructor, and indeed I think there is a place for such feedback.

But notice that my main claim here is a narrow one, namely, that using SETs *as a measure of teacher effectiveness* is morally wrong because it imposes an unfair risk of harm on instructors. Many supervisors at least act as if—and perhaps believe that—SETs measure teacher effectiveness. Second, if the goal were to measure student satisfaction, it would be more appropriate to ask questions such as "How much did you enjoy the course?" rather than "How much did you learn in this course?" Third, if supervisors use SETs to only measure student satisfaction, they lack an indicator that measures teaching effectiveness. It is implausible that higher education should primarily be about student satisfaction rather than learning.

Finally, what do the data tell us about the probability that SETs accurately measure an individual instructor's teaching effectiveness? Even using the highest correlation found in the literature, 0.4,[35] which was later revealed to be an overestimate,[36] Esarey and Valdes calculate that "over one quarter of faculty with SET scores at or below the 20th percentile are actually better at teaching than the median faculty member in our simulation. Even those with exceptionally high SET scores can be poor teachers: nearly 19% of those with SET scores above the 95th percentile are no better than the median professor at teaching."[37]

These studies provide reasons to think that SETs inadequately measure teaching effectiveness, especially when they compare teachers who use active learning methods of instruction with those who do not. The evidence suggests that part

of the issue with the construct validity of SETs is that students are not good at assessing whether they have learned something.[38] Thus, it is likely that SETs violate the first and second duties of university supervisors, namely, that the indicators supervisors use to evaluate employees have a reasonable likelihood of actually measuring the relevant concepts and do not incentivize wrongdoing. An additional problem of any indicator with a low or no construct validity is that it risks wrongfully harming instructors who do not deserve a low teaching evaluation. Another way supervisors can violate a duty is by using indicators that risk wrongfully discriminating against members of disadvantaged groups and not correcting for that discrimination. I turn now to that.

Evidence That SETs Are Likely to Perpetuate Discrimination

The evidence is mixed, but there are observational and experimental empirical studies that suggest that students tend to exhibit gender, racial, and cultural bias in teaching evaluations.[39] In a clever study, researchers took advantage of a class that was taught exclusively online in order to test gender discrimination.[40] Instructors with gender-identifiable names taught sections using their real names and sections where they switched names to that typical of another gender (a female using a male name, and vice versa). This directly controls for teaching ability because the same person is teaching, using the same teaching techniques, at nearly an identical time. Across a range of indicators, students rated the instructor with the traditionally female name lower than the instructor with the traditionally male name. Overall, students rated the perceived female professor 0.61 out of five points lower than her perceived male counterpart.[41] This study also has implications for the construct validity of teaching evaluations. If the students were learning the same amount from the same teacher no matter the name that he or she used, a reliable indicator would suggest equivalent student evaluations, on average. Similarly structured studies have had similar findings.[42] In other online courses where students knew the gender of their professors, the female instructor had lower SETs than her male counterpart.[43]

Are the substantive effect sizes large enough to meaningfully impact employment assessment? Several of the findings, combined with typical assessment methods, suggest they are. At many other universities, there are likely explicit or implicit targets. Presumably, if professors score at or above that level, they will achieve at least a satisfactory annual review score on that metric. If they score below that metric, they are at risk of receiving an unsatisfactory rating. The use of SETs can also have compound effects. To receive a permanent contract in The Netherlands, for example, universities generally require professors to acquire a teaching certificate (*Basis Kwalificatie Onderwijs* [BKO]) that is based in part on SETs. In sum, there is sufficient evidence to conclude that given a preponderance of evidence standard suggests that SETs likely perpetuate wrongful discrimination.

Given the evidence I presented above, using SETs in performance evaluations as indicators of teaching effectiveness likely violates all three duties of an employer. If an employer imposes a wrongful risk of harm if they violate any one of the first three duties, then they also do so if they violate all three. The evidence suggests that when supervisors use SETs as measures of teaching effectiveness in performance evaluations, they likely violate all three duties. Thus, in fact, many institutions that choose to use SETs in performance evaluations of teaching effectiveness harm many employees. A straightforward way to explain what is wrong with any of these is that they individually and collectively impose a wrongful risk of harm.

Possible Solutions

There are two types of solutions to the problems raised in this article: those that address the construct validity issue and those that attempt to correct for biases present in SETs without changing the fundamental means of instructor assessment. I discuss both because the second is useful in the near term given the high probability that many universities will continue to use SETs for the foreseeable future, whereas the former should be a more fundamental goal of universities and other instructional institutions.

One option would be for university supervisors to not use SETs in teacher assessments. It is better to use no metric than a flawed one, especially if those flaws are likely to wrongfully harm innocent people. In many universities, this seems unrealistic in the near future because university administrators are often resistant to dropping the use of SETs as performance metrics, even if these metrics are more unfair than no metric at all. Student evaluations of teachers provide a false veneer of fairness, objectivity, and accountability because of their face validity. They can be produced to defend personnel decisions. Universities can present them to review committees, bosses, students, and potential students' parents to show how impressive or problematic an instructor is.

A flawed solution would be to normalize teaching evaluation scores across genders and other factors of known biases. For instance, suppose that, on average, male teachers get 10 percent higher teaching evaluations than do female instructors. A simple solution would be to make comparisons fair by increasing female professors' scores by 10 percent. This rough measure should likely be adjusted given the characteristics of the instructor in question. For instance, junior women,[44] and women for whom English is not their first language,[45] score worse on SETs than similarly situated men. Another imperfect solution is to insert language into teaching evaluations that a randomized trial finds can reduce biases.[46] Or universities could use a numeric rating out of six instead of ten because there is some evidence that this reduces bias against female instructors.[47] These are poor solutions because they do not address the problem of the construct validity of SETs.

To address the construct validity issue, however, requires developing and using indicators that would meet the duty to use only indicators that can reasonably reliably measure instructors' teaching skills based on their merits. This would entail developing indicators that can measure student learning. First, to do this accurately, a baseline measure would be necessary to see how much students know before and after a given course, semester, year, or academic degree. This could include a test given before and after a course to measure how much students learn. To my knowledge, almost no university uses such techniques, likely because they are costly, and few people have incentives to spend political capital on such reforms. Second, we should think carefully about what we want to teach. We likely want to teach a range of skills and content knowledge. One widely agreed-upon skill is critical thinking. Of course, this is not the only skill or knowledge we hope students will learn during a program. Tests exist to measure critical thinking skills. One is Collegiate Learning Assessment (CLA), as discussed by Arum and Roksa.[48] A simple way to assess a university's program or year of study is to give students the CLA at the start and end of various academic programs to measure improvements in critical thinking and other skills. Minerva University did just that to measure how much critical thinking improved over the freshman year.[49] This is only one means to assess one aspect of what students should learn. A well-known problem with any such test is that it could incentivize instructors to "teach to the test" to the detriment of other important skills. Of course, if the test measures what we want students to learn, this is not a problem. A more comprehensive approach would be likely to develop specific means to assess student achievement and teacher effectiveness. Any potential reforms should be weighed against additional administrative problems that may themselves risk placing unfair burdens on some staff members.[50]

If one central goal is to improve student learning–as it surely should be–one evidence-based reform that supervisors should implement is to require instructors to use active learning techniques. Providing instructors with evidence-based instruction and requiring us to adopt these techniques are likely some of the best ways to translate the evidence-based means of learning into practice.[51] Currently, many instructors are not given any evidence-based instruction on how to teach or are given only a brief course at the start of their graduate career. Rather than using SETs in personnel assessments, supervisors and administrators could assess to what extent instructors adhere to the science of learning in their course design and implementation. The specifics of how this would be developed should probably be left to each particular program, but the broad ideas are well-understood: Does the professor only lecture or does she use active learning methods? Does she only require one final exam or are quizzes spaced out over the semester? A simple way to do this would be to review syllabi to see whether, say, there is only a final exam or paper, or whether there are quizzes given throughout a semester. Having some such standards could potentially reduce (unconscious) bias in evaluation methods.

Any such reform should avoid overburdening individual instructors. If universities were serious about this, they could devote resources to assisting professors in both studying the science of learning and adapting it to their courses to lighten the burden. This may have added benefit of achieving a version of a learning-based Rawlsian difference principle for those students most at risk of poor performance, as there is evidence that using active learning helps prevent the lowest-performing students from failing[52] and Pareto optimality for students.

Finally, perhaps program assessments rather than individual assessments should become standard. Instead of asking how effective teacher X is in class Y, such a shift in evaluation could allow broader questions that could improve student learning outcomes more than evaluations of each teacher for each class. For instance, even the best teachers will probably not help students put information, techniques, and critical thinking skills into long-term memory if the information happens to be taught at the end of a course, and students are not tested or required to use that information again in their studies. This is because to put ideas and facts into long-term memory, evidence suggests that effortful recall spaced over time is essential.[53] By integrating larger goals into programmatic assessments, however, learning architects could ensure that important concepts, facts, methods, and so on are repeatedly tested and used throughout a program. Programmatic assessments could begin with questions such as the following: What should students know when they graduate? Are current programs ideally structured to maximize student learning? How might the science of learning inform improvements to courses, programs, and majors? While this will not likely satisfy administrators and supervisors who want or are under pressure to have individual instructors rated on their teaching abilities, it might help students learn the most. In sum, these are not meant as definitive alternatives that should immediately replace SETs. Rather, they are ideas that deserve further research. I hope they contribute to a debate about how ends and means of education can be better achieved and instructors can be more fairly assessed than they are by the predominant current practices.

Conclusion

I have argued that imposing a wrongful risk of harm on instructors captures a large part of why using SETs in personnel assessments is problematic. I built this claim on several relatively uncontroversial premises regarding the moral duties that university supervisors have toward their instructors and that professors owe their students. These include supervisors treating teachers fairly, not incentivizing wrongdoing, and not using performance indicators that are likely to perpetuate discrimination, and a duty instructors owe their students to use the teaching methods most likely to result in students learning as much as possible. Student evaluations of teachers are problematic because it is likely they do not reliably measure student learning, they probably incentivize instructors to not use the most effective teaching methods, and

it is likely that they perpetuate discrimination against individuals who are members of certain disadvantaged groups. Therefore, there is a strong case for abolishing SETs as measures of teaching effectiveness and in instructor assessments.

Should SETs be abandoned altogether? Perhaps they should not be abolished for several reasons. One is that some indicators (other than teaching ability) may have adequate construct validity as measured by self-reporting. For instance, whether students feel like an instructor treats them with respect is at least partially subjective. Another question that might be usefully assessed by self-reporting is to what extent students feel like instructors are available for help outside of class. This could be assessed objectively as well, through such measures as the number of hours of office hours an instructor holds during a week. Students may exhibit biases in this self-reported measure, too, however, so investigators should be cautious about how much weight to put on such evaluations. Another way student feedback may be useful is in providing feedback to instructors, for instance, on workload, suggestions for other topic areas, and so on. These other uses of student feedback do nothing to undermine the main argument of this paper that there are robust reasons grounded in commonsense principles and strong evidence to suggest that SETs should not be used as they currently are in many institutions of higher education. The use of SETs as measures of teaching effectiveness imposes an unjust risk of harm on all instructors, but the risk is not evenly distributed. Some, especially minority instructors and those in precarious, teaching focused positions, are likely to have a higher risk of harm than others. This makes the use of SETs as measures of teaching effectiveness in personnel assessments especially wrong.

Leiden University
Eamon Aloyo
Turfmarkt 99
2511 DP The Hague
Netherlands
e.t.aloyo@fgga.leidenuniv.nl

NOTES

I thank the anonymous reviewer of this journal, Lydie Cabane, Honorata Mazepus, Vanessa Newby, Andrei Poama, and audiences at Leiden University's Institute Security and Global Affairs seminar, and the Disputationes Quadrangulae 2020 (including Wouter Kalf, Bruno Verbeek, and Axel Gosseries) for helpful feedback on earlier versions of this paper.

1. Heffernan, "Sexism, Racism," 1.
2. Brennan and Magness, *Cracks in the Ivory Tower*, chap. 4.
3. Brennan and Magness, *Cracks in the Ivory Tower*, 90, 95–100.
4. Brennan and Magness, *Cracks in the Ivory Tower*, 88.

5. Brennan and Magness, *Cracks in the Ivory Tower*, 88.

6. Brennan and Magness, *Cracks in the Ivory Tower*, 88.

7. Brennan and Magness, *Cracks in the Ivory Tower*, chap. 4.

8. Mill, *On Liberty*, § IV, 3.

9. Oberdiek, *Imposing Risk*, 15–17; Hansson, *Ethics of Risk*, 8–9.

10. Audi, Business Ethics, 75.

11. Brennan and Magness, *Cracks in the Ivory Tower*, 88–89.

12. Brennan and Magness, *Cracks in the Ivory Tower*, 93.

13. Brennan and Magness, *Cracks in the Ivory Tower*, 87.

14. Toshkov, *Research Design*, 173.

15. Lippert-Rasmussen, "Badness of Discrimination," 168.

16. Lippert-Rasmussen, "Badness of Discrimination," 174.

17. Segall, "What's So Bad."

18. Other factors besides a teacher's ability likely matter for student outcomes. For instance, there is some evidence from the United States that if a student and teacher share the same race, students perform better. See Dee ("Teachers, Race"); Dee ("Teacher Like Me"); Redding ("Teacher Like Me").

19. Parr, "Revisiting Harmless Discrimination," 1536.

20. McIntire and Keller, "Demand for Money."

21. Goodin, "Duty to Let Others."

22. Brown, Roediger, and McDaniel, *Make It Stick*; Deslauriers, Schelew, and Wieman, "Improved Learning"; Freeman et al., "Active Learning"; Haak et al., "Increased Structure."

23. Arum and Roksa, *Academically Adrift*, 36.

24. Arum and Roksa, *Academically Adrift*, 35–36.

25. Stroebe, "Why Good Teaching Evaluations"; "Student Evaluations."

26. Flaherty, "Teaching Eval Shake-Up."

27. Chetty et al., "How Does Your Kindergarten"; Staiger and Rockoff, "Searching for Effective Teachers."

28. Hessler et al., "Availability of Cookies"; Felton, Mitchell, and Stinson, "Web-Based Student Evaluations"; Riniolo et al., "Hot or Not."

29. Clayson, "Student Evaluations"; Onwuegbuzie, Daniel, and Collins, "Meta-Validation Model"; Spooren, Brockx, and Mortelmans, "On the Validity"; Uttl, White, and Gonzalez, "Meta-Analysis."

30. Uttl, White, and Gonzalez, "Meta-Analysis," 40.

31. Deslauriers et al., "Measuring Actual Learning"; Walker et al., "Delicate Balance."

32. Freeman et al., "Active Learning Increases"; Theobald et al., "Active Learning Narrows."

33. Brennan and Magness, *Cracks in the Ivory Tower*, 100–02.

34. Hirschman, Exit, Voice, and Loyalty.

35. Cohen, "Student Ratings."

36. Uttl, White, and Gonzalez, "Meta-Analysis."

37. Esarey and Valdes, "Unbiased, Reliable," 2.

38. Brown, Roediger, and McDaniel, *Make It Stick*, chap. 5.

39. Boring, "Gender Biases"; Drake, Auletto, and Cowen, "Grading Teachers"; Fan et al., "Gender and Cultural Bias"; Mengel, Sauermann, and Zölitz, "Gender Bias"; Miles and House, "Tail Wagging the Dog"; Reid, "Role of Perceived Race"; Heffernan, "Sexism, Racism"; Kreitzer and Sweet-Cushman, "Evaluating Student Evaluations."

40. MacNell, Driscoll, and Hunt, "What's in a Name."

41. Mitchell and Martin, "Gender Bias."

42. Chávez and Mitchell, "Exploring Bias."

43. Mitchell and Martin, "Gender Bias."

44. Mengel, Sauermann, and Zölitz, "Gender Bias."

45. Fan et al., "Gender and Cultural Bias."

46. Peterson et al., "Mitigating Gender Bias."

47. Rivera and Tilcsik, "Scaling Down Inequality."

48. Arum and Roksa, Academically Adrift.

49. Kosslyn, "Minerva Delivers."

50. Thanks to an anonymous reviewer for raising this important point.

51. Brown, Roediger, and McDaniel, *Make It Stick*.

52. Freeman et al., "Active Learning Increases"; Walker et al., "Delicate Balance"; Theobald et al., "Active Learning Narrows."

53. Karpicke and Roediger, "Expanding Retrieval Practice"; Karpicke and Bauernschmidt, "Spaced Retrieval."

REFERENCES

Arum, Richard, and Josipa Roksa. *Academically Adrift: Limited Learning on College Campuses*. 1st ed. Chicago: University of Chicago Press, 2011.

Audi, Robert. *Business Ethics and Ethical Business*. 1st ed. New York: Oxford University Press, 2008.

Boring, Anne. "Gender Biases in Student Evaluations of Teaching." *Journal of Public Economics* 145 (2017): 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006.

Brennan, Jason, and Phillip Magness. *Cracks in the Ivory Tower: The Moral Mess of Higher Education*. 1st ed. New York: Oxford University Press, 2019.

Brown, Peter C., Henry L. Roediger, and Mark A. McDaniel. *Make It Stick: The Science of Successful Learning*. Cambridge, MA: Belknap Press, 2014.

Chávez, Kerry, and Kristina M. W. Mitchell. "Exploring Bias in Student Evaluations: Gender, Race, and Ethnicity." *PS: Political Science & Politics* 53, no. 2 (2020): 270–74. https://doi.org/10.1017/S1049096519001744.

Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project Star." *Quarterly Journal of Economics* 126, no. 4 (2011): 1593–660. https://doi.org/10.1093/qje/qjr041.

Clayson, Dennis E. "Student Evaluations of Teaching: Are They Related to What Students Learn?: A Meta-Analysis and Review of the Literature." *Journal of Marketing Education* 31, no. 1 (2009): 16–30. https://doi.org/10.1177/0273475308324086.

Cohen, Peter A. "Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies." *Review of Educational Research* 51, no. 3 (1981): 281–309. https://doi.org/10.3102/00346543051003281.

Dee, Thomas S. "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?" *American Economic Review* 95, no. 2 (2005): 158–65.

———. "Teachers, Race, and Student Achievement in a Randomized Experiment." *Review of Economics and Statistics* 86, no. 1 (2004): 195–210. https://doi.org/10.1162/003465304323023750.

Deslauriers, Louis, Logan S. McCarty, Kelly Miller, Kristina Callaghan, and Greg Kestin. "Measuring Actual Learning versus Feeling of Learning in Response to Being Actively Engaged in the Classroom." *Proceedings of the National Academy of Sciences* 116, no. 39 (2019): 19251–57. https://doi.org/10.1073/pnas.1821936116.

Deslauriers, Louis, Ellen Schelew, and Carl Wieman. "Improved Learning in a Large-Enrollment Physics Class." *Science* 332, no. 6031 (2011): 862–64. https://doi.org/10.1126/science.1201783.

Drake, Steven, Amy Auletto, and Joshua M. Cowen. "Grading Teachers: Race and Gender Differences in Low Evaluation Ratings and Teacher Employment Outcomes." *American Educational Research Journal* 56, no. 5 (2019): 1800–33. https://doi.org/10.3102/0002831219835776.

Esarey, Justin, and Natalie Valdes. "Unbiased, Reliable, and Valid Student Evaluations Can Still Be Unfair." *Assessment & Evaluation in Higher Education*, February 20, 2020. https://www.tandfonline.com/doi/abs/10.1080/02602938.2020.1724875.

Fan, Yanan, Laura J. Shepherd, Eve Slavich, David Waters, Meryl Stone, Rachel Abel, and Emma L. Johnston. "Gender and Cultural Bias in Student Evaluations: Why Representation Matters." *PLoS ONE* 14, no. 2 (2019): e0209749. https://doi.org/10.1371/journal.pone.0209749.

Felton, James, John Mitchell, and Michael Stinson. "Web-Based Student Evaluations of Professors: The Relations between Perceived Quality, Easiness, and Sexiness." *Assessment & Evaluation in Higher Education* 29, no. 1 (2004): 91–108. https://doi.org/10.1080/0260293032000158180.

Flaherty, Colleen. "Teaching Eval Shake-Up." *Inside Higher Ed*, May 22, 2018. https://www.insidehighered.com/news/2018/05/22/most-institutions-say-they-value-teaching-how-they-assess-it-tells-different-story.

Freeman, Scott, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. "Active Learning Increases Student Performance

in Science, Engineering, and Mathematics." *Proceedings of the National Academy of Sciences* 111, no. 23 (2014): 8410–15. https://doi.org/10.1073/pnas.1319030111.

Goodin, Robert E. "The Duty to Let Others Do Their Duty." *Journal of Ethics* 24, no. 1 (2020): 1–10. https://doi.org/10.1007/s10892-019-09318-x.

Haak, David C., Janneke Hille Ris Lambers, Emile Pitre, and Scott Freeman. "Increased Structure and Active Learning Reduce the Achievement Gap in Introductory Biology." *Science* 332, no. 6034 (2011): 1213–16. https://doi.org/10.1126/science.1204820.

Hansson, Sven Ove. *The Ethics of Risk: Ethical Analysis in an Uncertain World*. 1st ed. Basingstoke, Hampshire, UK: Palgrave Macmillan, 2013.

Heffernan, Troy. "Sexism, Racism, Prejudice, and Bias: A Literature Review and Synthesis of Research Surrounding Student Evaluations of Courses and Teaching." *Assessment & Evaluation in Higher Education* 47, no. 2 (2021): 1–11. https://doi.org/10.1080/02602938.2021.1888075.

Hessler, Michael, Daniel M. Pöpping, Hanna Hollstein, Hendrik Ohlenburg, Philip H. Arnemann, Christina Massoth, Laura M. Seidel, Alexander Zarbock, and Manuel Wenk. "Availability of Cookies during an Academic Course Session Affects Evaluation of Teaching." *Medical Education* 52, no. 10 (2018): 1064–72. https://doi.org/10.1111/medu.13627.

Hirschman, Albert O. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Cambridge, MA: Harvard University Press, 1970.

Karpicke, Jeffrey D., and Althea Bauernschmidt. "Spaced Retrieval: Absolute Spacing Enhances Learning Regardless of Relative Spacing." *Journal of Experimental Psychology. Learning, Memory, and Cognition* 37, no. 5 (2011): 1250–57. https://doi.org/10.1037/a0023436.

Karpicke, Jeffrey D., and Henry L. Roediger. "Expanding Retrieval Practice Promotes Short-Term Retention, but Equally Spaced Retrieval Enhances Long-Term Retention." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33, no. 4 (2007): 704–19. https://doi.org/10.1037/0278-7393.33.4.704.

Kosslyn, Stephen M. "Minerva Delivers More Effective Learning: Test Results Prove It." Blog post on Medium.com, Minerva University, October 10, 2017. https://medium.com/minerva-university/minerva-delivers-more-effective-learning-test-results-prove-it-dfdbec6e04a6.

Kreitzer, Rebecca J., and Jennie Sweet-Cushman. "Evaluating Student Evaluations of Teaching: A Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform." *Journal of Academic Ethics* 20, no. 1 (2022): 73–84. https://doi.org/10.1007/s10805-021-09400-w.

Lippert-Rasmussen, Kasper. "The Badness of Discrimination." *Ethical Theory and Moral Practice* 9, no. 2 (2006): 167–85. https://doi.org/10.1007/s10677-006-9014-x.

MacNell, Lillian, Adam Driscoll, and Andrea N. Hunt. "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching." *Innovative Higher Education* 40, no. 4 (2015): 291–303. https://doi.org/10.1007/s10755-014-9313-4.

McIntire, Mike, and Michael H. Keller. "The Demand for Money behind Many Police Traffic Stops." *New York Times*, October 31, 2021. https://www.nytimes.com/2021/10/31/us/police-ticket-quotas-money-funding.html.

Mengel, Friederike, Jan Sauermann, and Ulf Zölitz. "Gender Bias in Teaching Evaluations." *Journal of the European Economic Association* 17, no. 2 (2019): 535–66. https://doi.org/10.1093/jeea/jvx057.

Miles, Patti, and Deanna House. "The Tail Wagging the Dog: An Overdue Examination of Student Teaching Evaluations." *International Journal of Higher Education* 4, no. 2 (2015): 116–26.

Mill, John Stuart. *On Liberty*. New York: Everyman's Library, 1859.
Mitchell, Kristina M. W., and Jonathan Martin. "Gender Bias in Student Evaluations." *PS: Political Science & Politics* 51, no. 3 (2018): 648–52. https://doi.org/10.1017/S1049096518000014.
Oberdiek, John. *Imposing Risk: A Normative Framework*. 1st ed. New York: Oxford University Press, 2017.
Onwuegbuzie, Anthony J., Larry G. Daniel, and Kathleen M. T. Collins. "A Meta-Validation Model for Assessing the Score-Validity of Student Teaching Evaluations." *Quality & Quantity* 43, no. 2 (2009): 197–209. https://doi.org/10.1007/s11135-007-9112-4.
Parr, Tom. "Revisiting Harmless Discrimination." *Philosophia* 47, no. 5 (2019): 1535–38. https://doi.org/10.1007/s11406-018-0052-0.
Peterson, David A. M., Lori A. Biederman, David Andersen, Tessa M. Ditonto, and Kevin Roe. "Mitigating Gender Bias in Student Evaluations of Teaching." *PLoS ONE* 14, no. 5 (2019). https://doi.org/10.1371/journal.pone.0216241.
Redding, Christopher. "A Teacher Like Me: A Review of the Effect of Student–Teacher Racial/Ethnic Matching on Teacher Perceptions of Students and Student Academic and Behavioral Outcomes." *Review of Educational Research* 89, no. 4 (2019): 499–535. https://doi.org/10.3102/0034654319853545.
Reid, Landon D. "The Role of Perceived Race and Gender in the Evaluation of College Teaching on RateMyProfessors.com." *Journal of Diversity in Higher Education* 3, no. 3 (2010): 137–52. https://doi.org/10.1037/a0019865.
Riniolo, Todd C., Katherine C. Johnson, Tracy R. Sherman, and Julie A. Misso. "Hot or Not: Do Professors Perceived as Physically Attractive Receive Higher Student Evaluations?" *Journal of General Psychology* 133, no. 1 (2006): 19–35. https://doi.org/10.3200/GENP.133.1.19-35.
Rivera, Lauren A., and András Tilcsik. "Scaling Down Inequality: Rating Scales, Gender Bias, and the Architecture of Evaluation." *American Sociological Review* 84, no. 2 (2019): 248–74. https://doi.org/10.1177/0003122419833601.
Segall, Shlomi. "What's So Bad about Discrimination?" *Utilitas* 24, no. 1 (2012): 82–100. https://doi.org/10.1017/S0953820811000379.
Spooren, Pieter, Bert Brockx, and Dimitri Mortelmans. "On the Validity of Student Evaluation of Teaching: The State of the Art." *Review of Educational Research* 83, no. 4 (2013): 598–642. https://doi.org/10.3102/0034654313496870.
Staiger, Douglas O., and Jonah E. Rockoff. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24, no. 3 (2010): 97–118. https://doi.org/10.1257/jep.24.3.97.
Stroebe, Wolfgang. "Student Evaluations of Teaching Encourages Poor Teaching and Contributes to Grade Inflation: A Theoretical and Empirical Analysis." *Basic and Applied Social Psychology* 42, no. 4 (2020): 276–94. https://doi.org/10.1080/01973533.2020.1756817.
———. "Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluations." *Perspectives on Psychological Science* 11, no. 6 (2016): 800–16. https://doi.org/10.1177/1745691616650284.
Theobald, Elli J., Mariah J. Hill, Elisa Tran, Sweta Agrawal, E. Nicole Arroyo, Shawn Behling, Nyasha Chambwe, et al. "Active Learning Narrows Achievement Gaps for Underrepresented Students in Undergraduate Science, Technology, Engineering, and Math." *Proceedings of the National Academy of Sciences* 117, no. 12 (2020): 6476–83. https://doi.org/10.1073/pnas.1916903117.

Toshkov, Dimiter. *Research Design in Political Science*. 1st ed. London: Palgrave, 2016.
Uttl, Bob, Carmela A. White, and Daniela Wong Gonzalez. "Meta-Analysis of Faculty's Teaching Effectiveness: Student Evaluation of Teaching Ratings and Student Learning Are Not Related." In "Evaluation of Teaching: Challenges and Promises," edited by Fadia Nasser-Abu Alhija. Special issue, *Studies in Educational Evaluation* 54 (2017): 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007.
Walker, J. D., Sehoya H. Cotner, Paul M. Baepler, and Mark D. Decker. "A Delicate Balance: Integrating Active Learning into a Large Lecture Course." *CBE Life Sciences Education* 7, no. 4 (2008): 361–67. https://doi.org/10.1187/cbe.08-02-0004.