April 1, 2021

TO: Jason Stajich, Chair
    Academic Senate

FROM: Philip Brisk, Chair
    BCOE Executive Committee

RE: Ad Hoc Committee Report on Evaluation of Teaching

Dear Jason,

The BCOE Executive Committee reviewed the Final Report of the Ad Hoc Committee on Evaluation of Faculty Teaching. While the Executive Committee appreciates the efforts of the Ad-Hoc Committee, the Report fails to substantially address the fact that student evaluations of teaching (SETs) are irrevocably flawed, and their use as evidence of teaching in the merit and promotion process incentivizes grade inflation and fan service over assessment-driven improvements to pedagogy.

I will begin this letter by quoting directly from an article entitled “Sexism, racism, prejudice, and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching” by Professor Troy Heffernan at La Trobe University, Melbourne, Australia, published on March 6, 2021, in the journal Assessment & Evaluation in Higher Education (DOI: 10.1080/02602938.2021.1888075). Citing prior work, which I will not reproduce here, the paper notes that (1) trends in abusive comments in SETs are increasing; and (2) prejudices in SET results and the way that they are used is leading to increasing issues relating to academic mental health and wellbeing. I will quote the conclusion directly, noting that the primary concern of this paper is not the ineffectiveness of SETs in general, but their impact in marginalized groups from a DEI perspective:

This paper has shown that no university, and indeed the higher education sector as a whole, can declare to be a gender equal employer or have an interest in growing a safe, inclusive and diverse workforce if they continue using SETs to evaluate course and teacher quality.

This paper provides an evidence base which can be used as part of the growing material and argument against the practice of collecting SET data. When SET data is known to be highly prejudiced against many groups, methods must be changed, and using SET data as a component of hiring, promotion and/or firing decisions must be seen as the blatantly discriminatory practice that it is.

The need for immediate policy changes is clear. Women and marginalised groups are losing jobs, failing to achieve promotion, and are being negatively impacted at every step where SETs are concerned, and will continue to be discriminated against every time SET data is collected until the practice is stopped. These practices not only harm the sector’s women and most underrepresented and vulnerable, it cannot be denied that SETs also actively contribute to further marginalising the groups universities declare to protect and value in their workforces.

The current practice of willfully including student perspectives in the evaluation of teaching using a flawed and biased measure does just as much disservice to the students as it does to the faculty. The BCOE Executive Committee values student feedback on teaching and strongly encourages the Senate to consider effective and unbiased mechanisms to incorporate student feedback into the Merit and Promotion process.
Having read the Ad Hoc Committee’s Report in detail, the BCOE Executive Committee feels that the majority of its suggestion represent incremental improvements in comparison to the status quo; to conserve space, the remainder of this letter will set forward some specific BCOE concerns about the adverse of impact of how SETs are presently used today, provide specific feedback on some of the suggestions in the Report, add a few unordered comments from various Committee members, and summarize some ideas put forward by Kathryn Hammar, the Student Representative on the BCOE Executive Committee.

The BCOE Executive Committee’s biggest overall concern has to do with the fact that many BCOE majors take the vast majority of their first- and second-year courses outside of BCOE in subjects such as English, Mathematics, Physics, Chemistry, and Biology. The SETs for the instructors who teach those courses are reviewed by members of those Departments and the respective Deans in CHASS and CNAS, and one BCOE Representative on CAP (who maintains confidentiality and provides feedback to no one). From the BCOE perspective, student “satisfaction” with their experience in these courses (as measured by SETs) is far less important than whether or not these courses have adequately prepared the students for success in upper-division engineering coursework for which these courses are prerequisites. Without loss of generality, an instructor who teaches an introductory course in calculus or physics who has low SETs via iEval is incentivized to improve those evaluations to optimize their likelihood of successful merit and promotion; this incentivization is completely decoupled from BCOE’s interests, which is to ensure that students are properly prepared for rigorous upper-division coursework. The fact that the merit and promotion process at UCR is in part tied to student opinion of lower-division preparatory coursework, but is wholly decoupled from any actual measure of the effectiveness of that preparation, is fundamentally flawed.

While Subcommittee A’s overall recommendation states in earnest that “there must be a mechanism for students to communicate their experience,” that fact that UCR exclusively measures student satisfaction but not learning outcomes and academic preparation, and bases merit and promotion decisions on this factor alone with respect to classroom teaching, irreparably degrades the institution. The reality of this farce is that BCOE students struggle in their courses and BCOE instructors suffer from lower SETs due to frustrated students who were poorly prepared by non-BCOE instructors who were incentivized to optimize their SET scores; in turn, BCOE instructors are incentivized to play the same game. SETs incentivize the wrong objectives, and the result is that they fail to serve both students and faculty well; moreover, the question of whether or not SETs serve the interests of prospective employers of UCR graduates is not considered.

The problems with SETs are specific examples of a much more general problem with these types of assessments:

**Campbell’s Law:** The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.

**Goodhart’s Law:** Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.

**Goodhart’s Law (as restated by Strathern):** When a measure becomes a target, it ceases to be a good measure.

The current merit and promotion process uses SETs as a measure of teaching excellence; however, it is clearly a target: “How can I improve my SET score?” is not the same question as “How can I improve my teaching effectiveness?” The dichotomy is even greater for underrepresented faculty: the gap between “How can I improve my SET score in the presence of bias?” and “How can I improve my teaching effectiveness?” may be insurmountable. Informing students about bias upfront and replacing
Another form of bias elucidated by Heffernan’s paper was the difference in evaluation scores between instructors who teach qualitative and quantitative subjects. The paper reported that faculty who teach quantitative subjects were 3.3 times more likely to score in a lower evaluation bracket than those being evaluated by students in qualitative subjects. Those teaching quantitative subjects were also 1.88 times more likely to “fail” their evaluations. This is not surprising, as quantitative subjects have a reputation for requiring more effort than qualitative subjects, and do require strong grounding in prerequisite for students to do well in exams. When students are unwilling or unable to work to overcome inadequate preparation (which may not be their fault) they blame the instructor. SETs are particularly trenchant if students have managed to obtain good grades in their prerequisites due to grade inflation without developing a thorough understanding of the material. Once again, the incentives do not align with optimizing teaching effectiveness.

Another concern is accusations of academic misconduct. The policy of SCAIP requires instructors to report misconduct within 30 days of discovering it; given that students accused of misconduct are unlikely to give positive SETs, this disincentivizes reporting misconduct and punishes instructors who do so, except when misconduct is discovered late in the quarter. This issue becomes more salient with the emergence of websites like Chegg and CourseHero and the fact that many students have established their own communities using services such as Discord outside of UCR’s purview. If academic misconduct is to be taken seriously, the merit and promotion system must protect faculty from adverse impacts when misconduct is reported, rather than creating a situation where they can explicitly be punished for doing their job properly.

Making matters worse, the merit and promotion process gives faculty little opportunity to respond to student reviews. There is insufficient space in the reasonably short self-statements that are used for merits; this is less of a concern for promotions, where the self-statements have no page limit. The BCOE Executive Committee supports efforts to institutionalize the use of self-reflective teaching statements, as suggested by the Academic Council Teaching Evaluation Task Force and the UC Centers for Teaching and Learning and suggested as a Phase I stopgap measure in the Ad Hoc Committee’s report.

With respect to the contents of the Task Force Report, the BCOE Executive Committee makes the following comments.

Subcommittee A’s overall recommendations:

- Subcommittee A’s general suggestions to incrementally modify the iEval process is insufficient and disappointing, given the outcomes of the 2020 UCR campus survey on iEval reported in Appendix B. Changing the questions in iEval is a superficial change, as it does not adequately address student learning, student preparation for subsequent classes in prerequisite chains, etc.

- The BCOE Executive Committee views favorably the suggestion that “additional forms of evidence of teaching evidence be widely adopted.” The report did not precisely clarify what other forms of evidence should be considered, although some ideas can be taken from the Academic Council Teaching Evaluation Task Force, but hopes that at the very least self-reflective teaching statements be included.

- The suggestion to maintain the five-point numerical scale but without comparison to Department and Campus means and medians is sensible; however, it is unclear what numerical value represents a threshold for acceptable teaching quality.

- The BCOE Executive Committee supports the nine specific recommendations for the iEval revision; however, there is some concern about Recommendation 5, which asks about Course Objectives. At present, Course Objectives are a best-practice, but are not institutionalized (although in BCOE, objectives are effectively required for ABET accreditation); however, the process of communicating
them to students has not been institutionalized. To the best of our understanding, it is recommended that instructors include the objectives on syllabi, but that may be insufficient when students are rapidly completing an iEval survey. It is also unclear if the learning objectives are specific to the course, or if different instructors can alter the learning objectives. To proceed in this fashion, some solution is needed to institutionalize them -- the most aggressive form of which might be imposing a Senate approval process (this is not an endorsement of the suggestion to do so).

• Subcommittee A’s report incorrectly labeled Associate Professor Suveen Mathaudhu as being a member of the Physics Department; he is a member of the Mechanical Engineering Department.

• Among the potential iEval questions suggested by Faculty that are listed in Appendix D, the BCOE Executive Committee strongly opposes the following:
  
  o “Did this course lead you to want to take more courses in this field of study?” This question makes sense for students who are admitted to Colleges, but not directly to majors, and take many general education courses in the first two years while choosing their major. It does not make sense for students in BCOE who, depending on their major, take the major of their courses in the Calculus, Physics, Chemistry, and/or Biology sequences, plus English 001A and 001B; and only take courses within their majors in upper division. Likewise, it is not a good question for upper-division courses that are not prerequisite for other courses in the major, and are likely to be taken by students in their Junior and Senior years.
  
  o “The instructor appeared to reflect and tried to adapt their teaching when things didn’t go as planned or when students were struggling.” This question assumes that problems are the norm; it is unfair to ask the question if things do in fact go as planned when a course is offered; it also assumes that students are qualified to assess when things do or do not go as planned, and whether or not they and their peers struggle. It also places responsibility for struggling students on the instructor, which may or may not be appropriate, depending on the context.
  
  o “Do you feel the course was a valuable experience in helping to develop your own thinking/direction/approach on the subject?” This type of question might be appropriate for a course in the social sciences or humanities, but would not make sense for many courses that BCOE students take (e.g., Calculus).

With respect to Subcommittee B’s report, the BCOE Executive Committee sees the potential for improvements in the current iEval process, but is not convinced that the fundamental flaws will be overcome. The UC Berkeley option to allow instructors to configure the evaluation questions is perhaps the most substantive improvement. Additionally, many of the specific items are quite thoughtful, such as Item 3: “How to incorporate for class size, level (lower/upper division, grad/undergrad, required/elective), type (lecture, online, hybrid, flipped), etc. in the interpretation of teaching evaluations?”

One concern is that Subcommittee B listed the following question as mandatory: “Overall Teaching Effectiveness (MANDATORY): Considering both the limitations and possibilities of the subject matter and the course, how would you rate the overall effectiveness of this instructor?” This is similar to Question #13 on the current iEval, which Subcommittee A identified as being among “those questions most prone to bias [that] should be excluded.”

Other potentially problematic questions include:
• CATEGORY 1: Instructor-specific question themes:
  o “Encouraging of participation/discussion: The instructor encouraged student questions and participation.” While this question may seem innocuous, it isn’t reasonable to expect participation and discussion in courses with 100+ students. Another concern is that many students do not attend due to various reasons, including jobs and perhaps misplaced priorities. The students actually do NOT want in-class participation to be incentivized or required, as their preference is not to attend. This question invites them to attack the instructor for what is generally considered to be good teaching practices.
  o “Useful/clear feedback on performance: The instructor provided clear constructive feedback.” UCR has grown its undergraduate student population rapidly over the past decade, and class sizes have increased dramatically. It is not possible for instructors to provide substantive feedback throughout the quarter when class sizes grow above 20-30 students. Reductions in certain types of teaching quality should be expected when class sizes grow and adequate resources, such as funding for TAs and readers, are scarce.

• Category II Course Specific Themes:
  o “Application & Skill Development: The course developed my abilities and skills for the subject.” The problem here is one that I alluded to earlier in this letter. A student may not know if they adequately developed their skills and abilities until a later course in the prerequisite chain. A student may feel confident in their abilities and skills at the completion of a Calculus course, but may find themselves woefully unprepared a few quarters later in an upper-division course on an engineering topic such as heat transfer.
  o “Theory/Content Knowledge: The course developed my ability to think critically about the subject.” This question is biased toward classes in the social sciences or humanities. Critical thinking in subjects like mathematics may not really come until graduate school. Engineering disciplines require students to learn things like “how to compute the integral of common types of functions” or “how to write a for-loop in Python,” not how to think critically about the greater ramifications (or lack thereof) when computing a Fourier Transform.

Subcommittee B’s report failed to consider many factors that contribute to the student experience in a class, and may impact SET scores, but are beyond the control of the instructor. Example is facilities (e.g., classroom, laboratories, performance spaces, etc.), which are briefly mentioned in Subcommittee A’s report, or issues like scheduling -- for example, students may be unhappy with a course that meets at 8am, without recognizing that this was not the instructor’s choice. Examples of questions that could lead to better evaluation of these factors could include:
  • Did the classroom environment affect learning?
  • Did the scheduling affect learning?
  • Adequacy of laboratories?

Somewhat more generally, the BCOE Executive Committee recommends that the iEval questions be vetted by faculty who regularly use surveys in their research.

The BCOE Executive Committee held a rather open-ended discussion on SETs that went far beyond the scope of the Ad-Hoc Committee Report. The Committee feels that many of these comments may be useful, they are reproduced here, in no specific order:
• iEval response rates were much higher when students were required to submit evaluations before seeing their grade. Now that they are optional, response rates are lower, and, similar to social media, tend to reflect the most polarized views. Opportunities for legitimate feedback are lost. There is general concern that they capture extreme opinion, not mean or median opinion, and that instructors are adjusting their teaching in responses to extremes. There could be severe adverse outcomes on student learning that are not measured as a consequence.

• A counterbalance to iEval, which are very short-term in terms of what they assess, could be continued engagement with alumni. Sometimes, the value of a class is not properly understood until years later, either in the workforce, graduate school, etc.

• BCOE is required to collect a monumental amount of assessment data as part of the ABET accreditation process. Moreover, BCOE programs are required to analyze and act upon that data. Instructors who assess and modify their courses through this continuous improvement process should be empowered to present evidence that they have done so in their merit and promotion files.

• Many students do not like some things that ABET requires, and may inadvertently blame the instructor on their SETs.

• It could make sense to drop the point/scoring system and focus exclusively on the written comments; coherent writing might yield a more holistic picture of an instructor.

• It is very challenging to tease out lies in the written comments (and presumably poor ratings given by the same students). This is especially true in cases where students have been accused of academic conduct violations, within 30 days of discovery, as required, but prior to the closing of the iEval period.

• Instructors could benefit from an iEval-based formative survey that is carried out midway through the quarter. This way instructors can show adjustment and can demonstrate impact with the present summative survey at the end of the class. These questions should focus both on the instructor and the student, for example:
  - What can the student do to make the class better?
  - What can the instructor do to make the class better?
  - A question about how the responses to the mid-quarter survey are used to adjust teaching could be asked at the end-of-quarter survey?

Lastly, the BCOE Executive Committee would like to highlight comments made by Kathryn Hammar, the Student Representative. Kathryn’s comments both highlight the weaknesses of the current approach to SETs, but also put forward some alternative mechanisms to incorporate student feedback and opinion in a more constructive manner.

• Many students do not know why they are asked to fill out SET; they know nothing of their institutional purpose.

• Many students do not care about improving the course unless they get a bad grade, and they just want to take it out on the professor.

• Making SETs optional creates bias in terms of which students choose to fill them out.
• Numerical responses that don’t clearly correlate to written comments that explain them are not particularly useful.

• Retention of learned information is challenging from one course to the next.

• Students will provide better qualified feedback if there is greater awareness and information about the role of iEval on campus. Student participation will increase if they believe that their feedback will improve future offerings of a course.

• Student leaders (e.g., Presidents and Vice Presidents of student organizations) can be an effective vehicle to promote change in the culture of how SETs are administered and interpreted. Student leaders could be trained to communicate the importance of critical feedback through SETs to their classmates. Examples where they can show evidence of change would increase student confidence that their opinions are taken seriously.

• Continuous feedback through the quarter, rather than just waiting until the end, would be beneficial.

• It is important to separate evaluation of teaching from the course and infrastructure, especially for majors that have many laboratory classes. For example, students may have a strong opinion about a laboratory manual, but may be unaware that it was not written by the instructor.

The BCOE Executive Committee feels that Kathryn’s comments about engaging student leaders in the process of evaluating teaching is a step in the right direction, although the precise form of engagement, time commitment, potential compensation, etc. are open questions that we did not discuss.
Sexism, racism, prejudice, and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching

Troy Heffernan

To cite this article: Troy Heffernan (2021): Sexism, racism, prejudice, and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2021.1888075

To link to this article: https://doi.org/10.1080/02602938.2021.1888075

Published online: 06 Mar 2021.
Sexism, racism, prejudice, and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching

Troy Heffernan
La Trobe University, Melbourne, Australia

ABSTRACT
This paper analyses the current research regarding student evaluations of courses and teaching. The article argues that student evaluations are influenced by racist, sexist and homophobic prejudices, and are biased against discipline and subject area. This paper’s findings are relevant to policymakers and academics as student evaluations are undertaken in over 16,000 higher education institutions at the end of each teaching period. The article’s purpose is to demonstrate to the higher education sector that the data informing student surveys is flawed and prejudiced against those being assessed. Evaluations have been shown to be heavily influenced by student demographics, the teaching academic’s culture and identity, and other aspects not associated with course quality or teaching effectiveness. Evaluations also include increasingly abusive comments which are mostly directed towards women and those from marginalised groups, and subsequently make student surveys a growing cause of stress and anxiety for these academics. Yet, student evaluations are used as a measure of performance and play a role in hiring, firing and promotional decisions. Student evaluations are openly prejudiced against the sector’s most underrepresented academics and they contribute to further marginalising the same groups universities declare to protect, value and are aiming to increase in their workforces.

Introduction
This paper is part of a research project that investigates questions concerning student evaluations of courses and teaching (SETs). The project examines the extent to which SET results provide objective assessments of the course, the academic teaching the course, how biased the information may be, and what groups are potentially disadvantaged by these processes. It also investigates what adjustments can be made to create a more equitable system. Ample evidence indicates that course evaluations (that is, evaluations of a course in terms of its content and outcomes) and teaching evaluations (surveys dedicated to evaluating the teaching academic’s performance), are both strongly correlated with the teaching academic’s demographics and other issues unrelated to the course or the academic’s performance (Boring, Ottoboni, and Stark 2016; Uttl and Smibert 2017). For this reason, course and teaching evaluations are treated as one within this paper (unless otherwise stated).
This paper’s collation and analysis of existing SET research makes several points clear: SETs are significantly biased due to the demographics of students completing them, and prejudice against the academic teaching the course, are dependent on subject areas, and are impacted on by myriad other aspects not connected to the teacher or course. Yet despite these clear prejudices and biases, SETs are used to gauge teaching quality and are a component in judging who is hired, who is let go and who is promoted. In addition to these discriminatory practices, the trend of abusive comments in SETs is increasing (Tucker 2014). These prejudices in SET results, and how the results are used, are subsequently leading to growing academic mental health and wellbeing issues that universities cannot ignore (Jordan 2011; Fan et al. 2019).

Thus, this paper provides much needed synthesis and analysis of the existing research for the benefit of academics in every field and discipline who are subjected to these practices. Cunningham-Nelson, Baktashmotlagh, and Boles (2019) have built upon the International Association of Universities ‘World List of Universities’ (2006) and estimated that globally the teaching staff of over 16,000 higher education institutions collect SETs at the end of each teaching period. The paper makes clear that while SETs are being used as an aid in gauging performance, women and marginalised groups are losing jobs, being promoted slower and/or less often, and are being negatively impacted at career progression junctures within the academy (Uttl and Smibert 2017). This paper aims to inform higher education institutions, policymakers and administrators of the sexist, racist, homophobic and other biases that underpin SET data as the evidence demonstrates the way institutions are complicit in prejudicial practices associated with SET data.

**Methods**

This paper analyses the literature and evident themes around SETs and the prejudices and biases that influence their results. The paper follows the methodology of a systematic analysis to provide a transparent account of how data was gathered and informs the audience of the protocols used to provide an account of the literature and themes from a breadth of sources (Macpherson and Holt 2007; McCrae and Purssell 2020).

**Data collection**

This paper’s literature review method used the following initial criteria. The search began by seeking out research published between 1990 and 2020. The period was selected to incorporate the growth of SET surveys to include how they are used today, how technology has changed how surveys are conducted, and how technology has contributed to new data analysis methods. The title, abstract and keyword search terms included all versions of ‘student evaluations’, ‘student evaluations of teaching’, ‘SETs’, ‘SECTs’, ‘course evaluations’ and ‘teacher evaluations’ in English language peer-reviewed articles and books from standard institutional databases including EBSCO, ProQuest and Web of Science.

The advantage of these search terms is that research surrounding these topics is not limited to publications explicitly dedicated to one field; in this case education researchers or education-based journals and books (Pittaway et al. 2004). As student evaluations impact on every university discipline, research has been conducted in essentially all disciplinary areas. These findings sometimes appear in journals relating specifically to the subject area (such as medicine or engineering) rather than journals focusing specifically on teaching that subject. These search terms thus resulted in literature beyond the usual scope of education journals and higher education researchers concerned with teaching practices.

The search resulted in 293 publications. The list was then refined to remove the papers addressing student surveys not connected to SETs (86), duplicate results (55), book reviews (9),
media articles (4) and non-English language publications (3). This resulted in 136 publications meeting the criteria. Analysis of these 136 articles then led to the identification of 47 articles that were not returned in the initial search. In most cases, the papers did not appear in the initial search because they were not publications dedicated to SETs; rather, the discussion around student surveys was carried out for the purpose of informing audiences concerned with course design or professional development. However, as they contributed to current scholarly thought around SETs and their impact, they were included in the following analysis.

Data analysis

After an initial reading of the final collection of material, themes were generated using Braun and Clarke’s (2006) method of thematic analysis. This analytic method provides a system by which patterns of meaning can be generated from qualitative data; in this case it was themes surrounding student evaluations of courses and teaching (Clarke and Braun 2017). This process identified 58 individual themes, which after further analysis were categorised into 35 sub-themes, before finally being refined into the five main themes discussed in this paper.

Findings

Evaluations technically work

A starting point of many publications examining SETs is an acknowledgement that evaluations work in the sense that they provide the university with data relating to course design, delivery of the course and teaching staff performance. The inherent problem with SETs, however, is that this data disguises the prejudices and biases underpinning the data being gathered, and subsequently in the results being produced (Marsh 2007; Osoian et al. 2010; Stark and Freishtat 2014). That SETs provide data that appears sound is arguably why institutions believe that evaluations are a measure of effectiveness (Osoian et al. 2010; Tucker 2014) and therefore play a role in academic hiring and promotions (Arthur 2009; Shah and Nair 2012; Boring, Ottoboni, and Stark 2016). At the same time, it has been noted that SET results have been used to aid in firing unproductive staff, or guiding staffing decisions during times of restructure (Jones, Gaffney-Rhys, and Jones 2014; Uttl and Smibert 2017).

Apparent data quality is also what leads institutions to use SETs as signifiers of teaching standards. Many institutions expect staff to achieve a certain SET result (e.g. 3.75 out of five or higher, or over 75% etc.) to be seen as fulfilling their duties. Stark and Freishtat (2014) also found that some universities use SETs to intentionally incite continuous cycles of competition amongst academics by making the acceptable result one that is above the cohort’s average. That is to say, no matter what the cohort’s results, half of the teaching staff will be below the average and susceptible to the negative repercussions of not meeting the target such as decreased promotion chances or leadership opportunities.

The issue many researchers examining SETs find is that looking at the data presented by compiling evaluation results appears to provide a somewhat objective picture of teacher and course success. However, what is rarely considered, or perhaps not seen, by universities and researchers praising SETs are the prejudices and conditions that shape the views that form the data (Marsh 2007; Osoian et al. 2010; Stark and Freishtat 2014).

Impact of students’ cultural and demographic backgrounds

The review found that students’ backgrounds and demographics in terms of gender, age, disciplinary area and study type can all impact on an academic’s SET results. That student demographics alone impact on SET results demonstrates just how flawed the system of evaluations
is, and to what extent results are determined by factors not related to course content or teaching quality (Rosen 2018).

Tucker’s (2014) study of 43,000 course evaluations found that across every discipline and course evaluated, and regardless of teaching staff, women students submitted SETs that were overall more favourable by two per cent. However, the standard level of positive SET score increased by as much as six per cent when completed by international students, students from older age groups, external students and students with higher grade point averages (GPAs). That student demographics alone can make a difference of between two and six per cent in SET results is a significant finding. However, it must be reiterated that this is just one of the many biases and prejudices that can accumulate to greatly disadvantage some groups more than others.

Variances in SET results being provided by students from different disciplinary areas have also been noted. Beran and Violato’s (2005) study of 370,000 evaluations and Centra’s (2009) study of 238,000 evaluations found that, across multiple universities and countries, academics teaching science-based subjects receive lower results than those teaching social science/humanities-based subjects. Uttl and Smibert (2017) study of 325,000 SETs divided subject areas into quantitative and qualitative subjects; those being subjects with assessments based on right/wrong assessment (such as correct calculations and formulas) and those with assessor judgements such as essays. They found that those teaching quantitative subjects were 3.3 times more likely to score in a lower evaluation bracket than those being evaluated by students in qualitative subjects. Those teaching quantitative subjects were also 1.88 times more likely to fail their evaluations. With quantitative subjects continually being found to be evaluated lower by students, Uttl and Smibert (2017, p. 8) concluded that:

Professors who teach quantitative vs. non-quantitative classes are not only likely to receive lower [SETs], but they are also at a substantially higher risk of being labelled unsatisfactory in teaching, and thus, more likely to be fired, not re-appointed, not promoted, not tenured, and denied merit pay.

Students are also influenced by factors not related directly to the course or teacher. Several studies have found high correlations between students’ grade expectations and the SET scores they deliver. These studies have examined expectations based on GPAs and mid-term results from institutions across several countries. Repeatedly, the findings are that students who are graded higher, or expect to gain a high grade if the SET is completed before results are released, provide higher scoring evaluations (Worthington 2002; Short et al. 2008; Stark and Freishtat 2014; Boring, Ottoboni, and Stark 2016). These findings also raise pedagogical concerns as these results provide motivation for academics to set easier assessment, or perhaps grade easier, to facilitate better SET results. Considering SET results are used to aid in hiring, firing and promotional decisions, it cannot be ignored that academics may be motivated to alter their assessment processes given that their livelihoods are at risk (Carter and Lara 2016; Bachan 2017).

Studies have also found that SET results are being driven by student biases irrelevant to course content and effectiveness. Benton, Cashin, and Manhattan (2012) found class size to be a major factor. Issues including classroom design, cleanliness of the university, quality of course websites, library services, food options available on campus, and difficulty in the admissions process (for first year students) have all been found to play a larger role in influencing SET results than teaching quality or course design (Osoian et al. 2010).

**Academic gender, ethnicity, sexual identity, and other demographics**

Arguments have also been made that an academic’s gender, ethnicity, language, perceived sexual identity, age or visible disabilities impacts on student evaluations (Valencia 2020). Such is the bias against gender and perceptions of ethnicity, sexuality, age and disability that, in 1993, Ambady and Rosenthal (1993) found that student reactions to a 30 second silent video of their
teacher played at the start of the semester correlated to the SET results the academic received. More recently, Boring, Ottoboni, and Stark (2016) study of 23,000 SETs, and Fan et al.'s (2019) study of 22,000 SETs, found that male students express a significant bias in favour of male academics.

Numerous studies have also found statistically significant differences between how gender influences academic evaluations. MacNell, Driscoll, and Hunt's (2015) work determined that women academics consistently receive lower scores relating to course design, clarity of assessment, class engagement, turnaround time of essays, and question response times regardless of their performance. Boring, Ottoboni, and Stark (2016) also found statistically significant examples of student expectations being amplified due to the academic's gender. Thus, they found that not only did SET results highly correlate with grade expectations, when grade expectations were met, male academics were rewarded with higher scores. When grade expectations were not met, the impact on evaluation scores was lower for male academics. To explore the extremes of gender prejudice in SETs, MacNell, Driscoll, and Hunt (2015) conducted a study that found that online classes led by male avatars (regardless of the academic's actual gender) received higher SET scores than those led by women avatars. Thus, students even perceiving the teaching academic to be a woman delivered lower scores.

A consistent theme within these studies is that gender, and even perceived gender, makes a difference to SET scores and is highly prejudiced against women. In Boring, Ottoboni, and Stark's (2016) summary of their study's findings and those of the existing studies they analysed, they declared that SETs are 'biased against female instructors by an amount that is large and statistically significant' (p. 1). Many of these studies find that these biases and prejudices result in large and significant variations in SET results, but this raises the question of what these differences can look like in practice. Fan et al.'s (2019) study of 22,000 SET results concluded that at the extreme, women academics are receiving SET scores 37 percentage points lower than male academics. This figure represents science subjects with high numbers of male students and being led by younger female academics (approximately under 35 years old). Boring, Ottoboni, and Stark's (2016) study found similar results and concluded that the biases at play were so great that more effective female academics are being placed in lower SET grading brackets than their less effective male counterparts.

In all studies relating to gender, the analyses indicate that the highest scores are awarded in subjects filled with young, white, male students being taught by white English first language speaking, able-bodied, male academics who are neither too young nor too old (approx. 35–50 years of age), and who the students believe are heterosexual. Most deviations from this scenario in terms of student and academic demographics equates to lower SET scores. These studies thus highlight that while, able-bodied, heterosexual, men of a certain age are not only the least affected, they benefit from the practice. When every demographic group who does not fit this image is significantly disadvantaged by SETs, these processes serve to further enhance the position of the already privileged.

Prejudice in SETs stemming from the academic's ethnicity is also a common finding in studies concerning evaluations and biases, as are issues around age, disability, sexual identity and appearance (Andersen and Miller 1997; Cramer and Alexitch 2000; Worthington 2002). A primary issue with these studies is that for large-scale quantitative research, the current state of higher education inclusion and diversity means results concerning marginalised groups are deemed too small to be considered a valid sample size (Hendrix 1998; Rubin 1998; Fan et al. 2019). However, smaller scale surveys and studies relying on qualitative methods consistently find prejudices against academics of colour (DiPietro and Faye 2005; Hamermesh and Parker 2005). Significant bias in SETs have been found to impact negatively on academics of colour (Hendrix 1998; Rubin 1998), and academics whose native language is not that of the university. In most studies this means the English language, though some studies have explored issues of SET prejudice in European non-English speaking countries and found similar results (DiPietro and
Fan et al.'s (2019) study also found that academics from diverse backgrounds or non-English first language backgrounds faced different levels of prejudice according to subject area; more liberal subjects were less prejudiced but were prejudiced nonetheless. Domestic students are also more likely to provide lower SET scores to academics from these groups than international students (Tucker 2014).

Providing further evidence of the significant prejudice women academics face in the SET process, even though all academics from ethnically diverse and marginalised backgrounds received lower SET scores than their white, English first language speaking colleagues, women from ethnically diverse backgrounds are graded more harshly than men from ethnically diverse backgrounds. At the extreme, Fan et al. (2019) found that in science faculties, a woman from a non-English speaking background was half as likely to receive the same SET result as a white English-speaking male.

**SET comments**

A key component of SETs in the literature that promotes their use is their anonymity. Universities argue that anonymity allows students to provide honest feedback without fear of retribution for speaking negatively against the teaching academic; the implication being that students may feel unable to make negative comments if they can be identified (Tucker 2014; Uttl and Smibert 2017). Studies have nonetheless determined that the issue with anonymous comments in SETs is that a portion of the comments are abusive, that the abuse is growing, and that the abuse is mostly directed towards race, gender, sexual identity, ethnicity, age and other marginalising characteristics.

The literature regarding comments in SETs is somewhat limited. Tucker’s (2014) study of 43,000 SETs suggested that only around one per cent of comments were abusive (though she noted sharp increases across studies, and hypothesised that the trend was growing rapidly). However, even if the figure of one per cent is accurate, it is imperative to point out that the one per cent is not distributed evenly across academic demographics. Additionally, to the person receiving the abusive comments, the emotional damage and stress is real, and the overall rate of abusive comments is irrelevant to their stress, anxiety or mental wellbeing (Tucker 2014).

As might be expected considering the clear trends within the findings of this review, a white male who is perceived to be heterosexual and is in the 35–50-year-old age group will, statistically speaking, receive few, if any, abusive comments. Abusive comments are mostly directed towards women and marginalised groups, focus on marginalising characteristics, and they are cumulative. For example, women receive abusive comments, and academics of colour receive abusive comments, thus, a woman of colour is more likely to receive abuse because of her gender and her skin colour (Oliver et al. 2008; Jordan 2011). One way to consider this finding is that some groups are so underrepresented in the sector that they do not constitute a valid sample size in large-scale studies. It is women and academics from these underrepresented groups that receive a majority of the abusive comments which makes the notion of ‘only one per cent of comments being abusive’ a highly distorted figure.

The result of SET comments is that they are a source of anxiety for academics. The process of SETs alone is cause for concern being that they are used for firing, hiring and promotion purposes (Jones, Gaffney-Rhys, and Jones 2014; Uttl and Smibert 2017). However, that comments for the sector’s women and marginalised groups are then likely to be at best unconstructive and unjustified, and at worst racist, sexist, homophobic or ageist (among other prejudices) is only a further cause for concern and mental distress for the academics receiving these comments (Jordan 2011; Tucker 2014).

Jones, Gaffney-Rhys, and Jones (2014) considered the legal implications of universities continuing to allow SETs to be collected when they are known to be a cause of distress. They discuss defamation and the university’s potential breaches of duty of care, and their suggestions
add further reason for SETs to be removed. Where comments are concerned, however, it is likely that if a student who could be identified provided racist, sexist, homophobic or other abusive feedback they would be removed from classes or face other consequences.

**What is known from previous SET research?**

For all the research and studies conducted around SETs, their conclusions largely all point to similar findings. A frequent conclusion across large-scale quantitative and smaller qualitative studies is that SETs rarely measure course or teacher quality or effectiveness, or if they do, these elements are significantly outweighed in the final results by other factors unrelated to course or teacher quality (Stark and Freishtat 2014; Tucker 2014; Boring, Ottoboni, and Stark 2016; Fan et al. 2019).

As has been discussed, SETs are heavily influenced by student demographics and subject area, but most studies argue that the two greatest influences are the academic’s gender and culture (Stark and Freishtat 2014; Boring, Ottoboni, and Stark 2016; Boring 2017; Fan et al. 2019). It is also critical to note that these researchers, and those of many more studies within the literature review, have suggested that it is impossible to account for these biases because from one class, subject, let alone one university, to another, the variances cannot be predicted or accounted for in the data.

Many researchers have not only questioned the use of SETs, but also the value of the content they provide. Cunningham-Nelson, Baktashmotlagh, and Boles (2019) make the point that even if SETs were not guided by biases and prejudice, the students completing the evaluation will not benefit from the process. The surveyed group will have moved on before changes can be made and the next cohort may value different attributes in the course or teacher. Researchers additionally argue that SETs are limited in what they provide because the qualities students place value on are already known (Sunindijo 2016). The existing literature declares that students want engaging lectures, assessment to be explained clearly and graded fairly, assessment returned in a timely manner and their questions answered promptly (Sunindijo 2016; Park and Dooris 2020). Vivanti, Haron, and Barnes (2014) came to similar conclusions, and add that students want guidance, and teaching staff who were interested in them and the subject they are teaching.

Similar arguments also appear in the literature alongside the conclusion that SETs are measuring customer satisfaction (Enache 2011; Osoian et al. 2010). These researchers acknowledge that the marketisation of the higher education system means customer satisfaction plays a valuable role in the modern university because enrolment numbers matter (Marginson 2013). However, these researchers are all highly critical that SETs are customer satisfaction surveys alleging to measure course outcomes and teacher effectiveness when many studies argue that this is untrue. These studies point out that customer satisfaction can be measured via methods that are not SETs, and suggest opportunities within the growing programs of universities embracing student voice and students as partners initiatives.

**Discussion**

This paper’s analysis of the existing literature makes it clear that SET results are strongly influenced by external factors unrelated to course content or teacher performance. In addition, these factors are frequently based on student demographics, and students’ biases and prejudices based on the teaching academic’s gender, sexuality, ethnicity, age or disability as well as other marginalising factors. Ultimately, this analysis raises the question of how can any university justify the continued use of SETs?

Official university responses to issues regarding SETs are rare. Tucker (2014) sought clarification of why a university continued using SETs when they were known to attract abusive
comments. The university's response was that the rate of abusive comments was too low to alter the existing procedures considering the value (they perceived) in the attained data. This response provides some insights into institutional administrative thinking. The first is that universities consider there to be an acceptable level of abuse that staff must endure. A factor here, however, is that Tucker collected and evaluated this information almost a decade ago. Considering the current focus on staff wellbeing and mental health (Henning 2018), it is possible that institutions today might have different views on academics receiving abuse. Universities might also be influenced by the knowledge that it is now known that abuse is directed heavily towards women and academics from marginalised groups. Thus, institutions continuing to conduct SETs are allowing the sector's most underrepresented and marginalised groups to be subjected to possible hate speech (Jones, Gaffney-Rhys, and Jones 2014; Uttl and Smibert 2017).

The second issue Tucker's (2014) research raised (as have other researchers in this literature review) is that universities believe the data they are collecting via SETs provides accurate assessments of course content and outcomes, and teacher quality and effectiveness. This paper's findings make it clear why universities might believe this to be the case. Methodologically, the process appears sound. Students provide anonymous feedback so as to not be concerned with potential repercussions from academics, and this data informs the faculty and administrators of students' perspectives concerning the course and teacher (Arthur 2009; Shah and Nair 2012; Jones, Gaffney-Rhys, and Jones 2014; Boring, Ottoboni, and Stark 2016). However, this review's findings also demonstrate that the methodology surrounding SETs is inherently flawed because the data being input into the system is influenced by biases and prejudices that are invisible in the data's results (Marsh 2007; Osoian et al. 2010; Stark and Freishtat 2014). These elements are also invisible because unless researchers specifically focus on biases in SETs, there is little reason to believe that the questions being asked will not be answered somewhat objectively across a class of respondents and provide an overall, somewhat accurate measure. However, this is not the case; this paper has shown that statistically significant biases in these data collection methods exist.

The groups most impacted by these prejudices are clear. This study demonstrates that at best SETs disadvantage women, and at worst, see women academics placed in untenable positions (MacNell, Driscoll, and Hunt 2015; Boring, Ottoboni, and Stark 2016). These results have also been theorised as a reason why women are underrepresented in both the professoriate and upper levels of university leadership (Fan et al. 2019).

The results are worse still when ethnicity, language, perceptions of sexuality or disability are considered. People from ethnic backgrounds, and/or who do not speak English as their first language, are receiving much lower SET results (Fan et al. 2019). Other researchers have examined disability, sexual identity, and cultural and linguistic diversity and found similar results in prejudice against these groups (Hendrix 1998; Rubin 1998; DiPietro and Faye 2005; Hamermesh and Parker 2005), though it cannot be ignored that researching prejudice against some of these groups is difficult because they are so underrepresented within the sector.

Researchers have also argued for decades that subject areas impact on SET results with the sciences being evaluated more harshly (Beran and Violato 2005; Centra 2009; Uttl and Smibert 2017). Similarly, that studies have routinely found high correlations between student grades and SET results makes it clear that student evaluations are being influenced by factors far exceeding the intended purpose of SET questions (Worthington 2002; Short et al. 2008; Stark and Freishtat 2014; Boring, Ottoboni, and Stark 2016). Studies have also determined that issues including classroom design, cleanliness of the university, quality of course websites, library services and food options available on campus all have a larger influence on SET results than practices concerning courses and teaching (Felton, Mitchell, and Stinson 2004; Felton et al. 2008; Osoian et al. 2010; Benton, Cashin, and Manhattan 2012).

In addition to the above findings, many researchers also argue that SETs are not needed because universities already know what students want from a course and teacher (Vivanti, Haron, and Barnes 2014; Sunindijo 2016; Park and Dooris 2020). These researchers claim that focus
groups or student interviews would be a more equitable way of gaining information and judging potential biases and prejudice. What these views make clear is that the growth in student voice, and students as partners initiatives, in recent years are well placed to provide discussions to improve learning opportunities rather than relying on the current prejudice practices.

Conclusion

This paper has shown that no university, and indeed the higher education sector as a whole, can declare to be a gender equal employer or have an interest in growing a safe, inclusive and diverse workforce if they continue using SETs to evaluate course and teacher quality.

This paper provides an evidence base which can be used as part of the growing material and argument against the practice of collecting SET data. When SET data is known to be highly prejudiced against many groups, methods must be changed, and using SET data as a component of hiring, promotion and/or firing decisions must be seen as the blatantly discriminatory practice that it is.

The need for immediate policy changes is clear. Women and marginalised groups are losing jobs, failing to achieve promotion, and are being negatively impacted at every step where SETs are concerned, and will continue to be discriminated against every time SET data is collected until the practice is stopped. These practices not only harm the sector’s women and most underrepresented and vulnerable, it cannot be denied that SETs also actively contribute to further marginalising the groups universities declare to protect and value in their workforces.

ORCID

Troy Heffernan http://orcid.org/0000-0002-8156-622X

References


