SAMT Report[1]

Student Assessment of Modules and Teaching: The Context

Over the past 25 years it has become increasingly evident that student evaluations of teaching are seriously flawed. They are flawed in two fundamental ways. First, high student evaluation scores appear to bear no meaningful relationship to learning outcomes.[2] In other words, the amount students actually learn plays no part in their assessment of the 'quality' of the teaching. Second, evaluations are deeply biased against lecturers who are minorities, women, disabled, have foreign accents, etc.[3] Taken together, the evidence shows that SAMTs are little more than popularity contests.

The higher education community ought to view this as extremely troubling. A university is fundamentally a place of learning. Hence it should promote and reward teaching that is genuinely high quality, not simply teaching that affords students an easy and enjoyable experience. An oncologist who is charming but cures no cancer is no use to the cancer patient. The flawed nature of SAMTs doesn't only affect students, however: the impact of SAMTs on educators is huge. Departments use them to ongoingly monitor their lecturers. SAMT scores are regularly used in making hiring, permanency, and promotion decisions. Though providing SAMT scores is not officially a requirement for these purposes except in the case of permanency applications, it is viewed as the norm and people are encouraged to supply them.

Even if the university were to ban the use of SAMTs in such cases, however, the unjustly lower scores and negative comments experienced by people from certain discriminated-against categories would continue to have a significant demoralizing impact, especially since many lecturers enter academia because they are passionate about good teaching. Indeed, some suggest that the ongoing lack of diversity in HE is partly due to SAMTS. Statistics obtained in 2013 by *Times Higher Education* found that only about 1 in 5 UK professors are female, with some universities having female professor populations as low as 8%. The demoralizing impact of poor SAMT results may prompt lecturers to allocate less time to their research in order to improve their teaching – with resulting damage to the career progression of already vulnerable populations (Mengel et al., 2017). Indeed, a male lecturer here at Essex offered the following anecdote as evidence of this disparity:

> "In the year just passed I taught on a new compulsory module with a number of teachers and the SAMT form allowed for rating of individual teachers as well as free form comments. The module was not popular and I was struck by the savagery of a couple of comments about a female colleague blaming her, quite unjustifiably, for the module's shortcomings."

---

[1] This report was inspired by a similar unpublished report written by Lara Owen for the Equity and Social Inclusion Committee (ESIC), Faculty of Business & Economics, Monash University entitled "Gender Bias in Student Evaluations" (2018).

[2] See, for example, Uttl et al., 2017; Galbraith, Merrill, & Kline, 2012; Carrell and West, 2010; Clayson, 2009; Schuck, Gordon, & Buchanan, 2008; Pounder, 2007; Zabaleta, 2007; Johnson, 2002; Abrami et al., 1982; Naftulin et al., 1973.

[3] See, for example, Flaherty 2018; Mitchell & Martin, 2018; Mengel et al., 2017; Boring, et al., 2016; MacNell et al., 2015; Tran, 2015; Bavishi et al., 2010; Reid, 2010; Boatright-Horowitz & Soeung, 2009; Youmans & Jee, 2007; Basow et al., 2006; Cambell et al., 2005; Basow et al., 2003.

Examples like this abound, to which many of us can all too easily attest. If we care about the lack of diversity in the lecturer population – as we purport to do – re-thinking our use of teaching evaluation is essential.

The University of Essex rightly prides itself on its leadership when it comes to considerations of equality and justice. Transforming the way that teaching is evaluated is an opportunity for Essex to demonstrate leadership again. In light of initiatives like Athena Swan it is clear that there is a growing recognition of the many ways in which women and minorities are dissuaded from staying in academia. It is time to take a leadership role in renouncing all practices that contribute to this chilling effect.

Change is Coming

There is a growing groundswell of recognition that universities cannot in good conscience continue to use SAMTs. For example, MacNell et al. (2015) argues that the evidence of bias in teaching evaluations of female lecturers indicates that the use of evaluations in employment decisions for women is discriminatory. Indeed, this issue was taken to arbitration in Canada (at Ryerson University in Toronto). The recent decision found that student evaluations of teaching cannot be used for promotion or tenure because they are known to be biased on the basis of protected classes: gender, race, age, accent (ethnicity, nat'l origin), etc. The language in the decision is quite strong:
https://www.canlii.org/en/on/onla/doc/2018/2018canlii58446/2018canlii58446.html.

Times Higher Education reports that in light of this arbitration, "Experts have predicted a global "sea change" away from the use of student evaluations to measure lecturers' suitability for promotion or tenure" (Bothwell 2018). Mitchell (2018) similarly argues that using such biased evaluative systems should be illegal on the basis of discrimination law: 'Our research shows they're biased against women. That means using them is illegal'; 'the use of student evaluations in hiring, promotion, and tenure decisions represents a discrimination issue. The [American] Equal Employment Opportunity Commission exists to enforce the laws that make it illegal to discriminate against a job applicant or employee based on sex. If the criteria for hiring and promoting faculty members is based on a metric that is inherently biased against women, is it not a form of discrimination?'

In light of these legal (and moral) worries, other universities have already begun to overhaul their teaching assessment systems to make them more fair: https://www.chronicle.com/article/A-University-Overhauled-Its/243803

Why Haven't Things Changed Yet?

As this report has been arguing (and will demonstrate in greater detail below), there is overwhelming evidence that student evaluation of teaching is a deeply flawed and unjust way to evaluate the quality of teaching. Why, then, does it continue to be used?

- Desire to give students a say in their education. 'Student voice' is and should be important to universities. But using SAMTs to capture this – and framing these

evaluations as an assessment of teaching quality – is predicated on a failure to appreciate the difference between student enjoyment and effective learning.

- Habit and ease. Not only is this the system that is already in place throughout the sector – thereby offering an air of legitimacy – but change is a hassle and alternatives might seem more difficult to administer.
- The mere fact that student evaluations are a metric gives them an un-earned air of rationality and scientific precision. Reducing the complexity of teaching to a single (meaningless) number also makes it possible easily to compare teacher performance.

What Does the Evidence Show?

What follows is a comprehensive overview of the evidence demonstrating the deeply problematic nature of SAMTs. As will become clear, merely 'tweaking' student evaluations – for example, by making them better designed surveys that don't trigger quite so much bias – will not succeed in removing their most troubling aspects. As Philip Stark, Associate Dean of Mathematical and Physical Sciences and a Professor of Statistics at the University of California, Berkeley has argued, "evaluations are biased against female instructors in so many ways that adjusting them for that bias is impossible" (Flaherty 2018).

What are the Biasing Factors?

The reason for Stark's pessimism about the possibility of salvaging SAMTs lies in the multiple arenas in which bias is operative:

1) ***student-related***: class attendance levels, students' degree of effort, expected and final grades, student gender, age, and pre-course interest and motivation all shape the way the student evaluates the perceived quality of the teaching. For example, Reisenwitz 2016 found that there are significant differences between those who complete evaluations and those who do not. See also Crumbley et al, 2001.
2) ***teacher-related***: similarly, the race, age, gender, reputation, research productivity, teaching experience, and personal traits unrelated to teaching (e.g. attractiveness) also bias students against some instructors more than others. (See Boysen, 2008; McNatt, 2010; Ambady & Rosenthal, 1993 and the section below entitled 'Bias against Instructors.'
3) ***course-related***: the class size, class attendance rate, class heterogeneity, course difficulty and workload, discipline, and level also play serious roles in student evaluation outcomes – providing further evidence that comparing different modules or lecturers on their basis is illegitimate (See Bassi et al. 2017).

Serious Flaws in the Structure of the Evaluations

Student evaluations of teaching simply assume that student are experts when it comes to what high-quality teaching is or how to recognize it (Newton 2007). Further, a host of studies have shown that student evaluations as they are currently structured are deeply flawed (Barrie, 2001; Barrie, Ginns, & Symons, 2008; Edström, 2008; Kolitch & Dean, 1999; Saroyan & Amundsen, 2001; Ray et al, 2018). Indeed, in 2016 Dr. Eric Jensen of Warwick – a sociologist and specialist in evaluation methodology – was invited by the Essex REO/Impact department to explain how to

design surveys for REF impact purposes. By coincidence the room had a dozen SAMT forms left on the tables. As an application, Dr. Jensen analyzed the form for the class, demonstrating how the Essex SAMT violates multiple rules of good survey design:

- The most common mistake was the lack of a 'neutral' answer (don't know, not applicable etc.).
- There were also questions that in effect asked multiple questions in one by using different adjectives.
- Many questions also used vague and highly subjective words like 'clear', 'good', and 'interesting' etc. without providing criteria for what those terms specify and which objective measures allow one to recognize instances of them.
- There was also variation in grading methodology from one question to the next. For example, for one question 1 = poor and 5 = excellent, but for other questions 5 = poor and 1 = excellent.

Relationship to Student Learning

Even if these flaws in survey design could be remedied, multiple studies demonstrate that high student evaluation scores do not necessarily track student learning. For example:

- Carrell and West (2010) concluded that high SET scores were actually associated with lower levels of deep learning. Students in sections run by older, stricter teachers did better in later courses than students taught first by more popular/lenient teachers.

- Naftulin et al. (1973). Investigators found that a professional actor who delivered a lecture on nonsense could extract higher evaluations from students than the experts.

- Abrami et al. (1982). Meta-analysis confirmed "instructor expressiveness" could drive student evaluations without improving student achievement.

- D.E. Clayson (2009). Found that once sample sizes were large enough, a teacher's evaluations were not linked to learning.

- Uttl et al (2017) performed a meta-analysis and discovered that the evidence strongly indicates that student evaluation of teaching is not related to student learning.

- V.E. Johnson (2002) found that evaluation scores correlate with the number of A's given, not the learning achieved.

- Galbraith, Merrill, & Kline, 2012; Pounder, 2007; Schuck, Gordon, & Buchanan, 2008; and Zabaleta, 2007 also support the findings that there are low or even no correlations between student evaluation scores and student learning.

Bias Against Instructors

The discriminatory and biased nature of student evaluations of teaching is also well-documented. A detailed list of further sources can be found below, but a sample of recent studies find the following:

- MacNell et al. (2015) used an online platform to disguise the gender of the teacher and identified significant gender bias in student evaluation. The instructor that students *thought* was a woman received significantly lower ratings on fairness, professionalism, respectfulness, enthusiasm, promptness, etc. The differences in 'promptness' scores were particularly striking, since work was marked and returned at the exact same time in both the 'male' and the 'female' led modules. Yet the lecturer students thought was male was given a 4.35 rating out of 5. The lecturer students thought was female received a 3.55.

- Reid (2010) engaged in a two-stage cluster analysis and found that student evaluations consistently deem white instructors best, with the worst-evaluated instructors more likely to be Black or Asian.

- Boring, et al (2016) also found that student evaluations of teaching are biased against female instructors by an amount that is large and statistically significant. This bias affects how students rate even putatively objective aspects of teaching, such as how promptly assignments are graded. The bias varies by discipline and by student gender, among other things. For example, male students are more likely to give lower scores to female lecturers than female students. The authors argue that it is not possible to adjust for the bias because it depends on so many factors. They also confirm other studies that find student evaluations to be more sensitive to students' gender bias and grade expectations than to teaching effectiveness. Indeed, gender biases can be significant enough to cause more effective instructors to get lower evaluation scores than less effective instructors. The authors also demonstrate that studies showing little evidence of bias are in fact flawed in their design: 'Not only are they observational studies rather than experiments, they ask the wrong question, namely, "do male and female instructors get similar SET [Student Evaluations of Teaching]?" A better question is, "would female instructors get higher SET but for the mere fact that they are women?" We can answer that question using these unique data sets: "yes."'

- Mitchell & Martin (2018) analyzed the language students use in evaluations and found that less respectful language is used towards female lecturers. **'**We found that a male professor was more likely to receive comments about his qualification and competence, and that refer to him as "professor."' 'We also found that a female professor was more likely to receive comments that mention her personality and her appearance, and that refer to her [simply] as a "teacher."' Mitchell & Martin also analyzed quantitative data, concluding that "Students appear to evaluate women poorly simply because they are women" (5).

- Dr. Ben Schmidt created a (truly eye-opening) interactive chart entitled 'Gendered Language in Teacher Reviews,' which lets you explore the words used to describe male

and female teachers in about 14 million reviews from RateMyProfessor.com. It can be found here: http://benschmidt.org/profgender/

Assessing Teaching Quality

In light of the profound flaws outlined above, using student evaluations to assess individual teachers/modules cannot be justified in an institution committed to both justice and high-quality teaching. This is especially the case when it comes to hiring, probation, and promotion.

UCU Essex seeks full partnership and representation in any consultations or reviews related to SAMT.

This is not to say that there shouldn't be quality-control measures in place, nor that we should ignore student experience. Both are essential. It is important to highlight, however, that there are already substantive teaching quality-control and student-voice procedures in place unrelated to SAMTs. These include: CADENZA, peer observation of teaching, student representatives, and Staff Student Liaison Committees, in which student representatives express their concerns about specific modules or the course more generally. Further, the elimination of SAMT would not foreclose the possibility of lecturers gathering module-specific feedback, but this would be neither mandatory nor a matter of public record.

By abolishing SAMTs in favour of such approaches to teaching quality-control and student voice, the University of Essex will be able to demonstrate its commitment to both high-quality teaching and fair treatment of its staff. It will be positioned to take a leadership role across the sector and serve as a beacon to students committed to both good teaching and principles of justice. We expect nothing less from this excellent institution.

Sources

Abrami, P., Leventhal, L., and Perry, R. (1982). "Educational Seduction," *Review of Educational Research* 52: 446-464.

Algozzine, B., Beattie, J., Bray, M., Flowers, C., Grets, J., Howley, L., Mohanty, G., and Spooner, F.(2004). "Student Evaluations of College Teaching: A Practice in Search of Principles." *College Teaching* 52: 134-41.

Barrie, S. C. (2001). "Reflections on Student Evaluation of Teaching: Alignment and Congruence in a Changing Context. In E. Santhanam (Ed.), *Student Feedback on Teaching: Reflections and Projections* (pp. 1-16). Perth: The University of Western Australia.

Basow, S., Phelan, J., and Capotosto, L. (2006). "Gender Patterns in College Students' Choices of their Best and Worst Professors." *Psychology of Women Quarterly* 30: 25-35.

Basow, S., Codos, S., and Martin, J. (2013). "The Effects of Professors' Race and Gender on Student Evaluations and Performance." *College Student Journal* 47: 352-63.

Bassi, F., Clerici, R., Aquario, D. (2017). "Students' evaluation of teaching at a large Italian university: validation of measurement scale" *Electronic Journal of Applied Statistical Analysis* 10: 93-117.

Bavishi, A., Hebl, M., and Madera., J. (2010). "The Effect of Professor Ethnic- ity and Gender on Student Evaluations: Judged Before Met." *Journal of Diversity in Higher Education* 3: 245-56.

Berk, R. (2005). "Survey of 12 Strategies to Measure Teaching Effectiveness." *International Journal of Teaching and Learning in Higher Education* 17: 48-62.

Boatright-Horowitz, S. and Soeung, S. (2009). "Teaching White Privilege to White Students Can Mean Saying Good-bye to Positive Student Evaluations." *American Psychologist* 64: 574-75.

Boring, A., Ottoboni, K, and Stark, P. (2016). "Student evaluations of teaching (mostly) do not measure teaching effectiveness" *Science Open Research* *https://www.scienceopen.com/document?vid=818d8ec0-5908-47d8-86b4-5dc38f04b23e*

Bothwell, Ellie. 2018. "'Tide turning' against using student evaluations to rate staff: Experts predict ripple effect from ruling against Canadian university" *Times Higher Education*, July 26, 2018. https://www.timeshighereducation.com/news/tide-turning-against-using-student-evaluations-rate-staff

Boysen, G. A. (2008). "Revenge and student evaluations of teaching." *Teaching of Psychology*, 35: 218-222. DOI: http://dx.doi.org/10.1080/00986280802181533

Braga, M., Paccagnella, M., and Pellizzari, M. (2014). "Evaluating Students' Evaluations of Professors." *Economics of Education Review* 41: 71- 88.

Campbell, H., Gerdes, K., and Steiner, S. (2005). " 'What's Looks Got to Do With It?' Instructor Appearance and Student Evaluations of Teaching." *Journal of Policy Analysis and Management* 24: 611-20.

Carrell, S. and West, J. (2010). "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors" *Journal of Political Economy* 118: 409-432.

Centra, J. (2003). "Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work?" *Research in Higher Education* 44: 495-518.

Clayson, D. E. (2009). "Student Evaluations of Teaching: Are They Related to What Students Learn? A Meta-Analysis and Review of the Literature" *Journal of Marketing Education.* 31: 16-30.

Crumbley, D. L., Henry, B., and Kratchman, S. (2001). "Students' perceptions of the evaluation of college teaching." *Quality Assurance in Education* 9: 197-207. DOI: http://dx.doi.org/10.1108/EUM0000000006158

Flaherty, C. (2018). "Teaching Eval Shake-Up" *Inside Higher Ed,* May 22, 2018.

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

Ginns, P., and Symons, R. (2008). *Student surveys on teaching and learning: Final report*. Sydney: Australian Teaching and Learning Council. http://www.itl.usyd.edu.au/cms/files/Student_Surveys_on_Teaching_and_Learning.pdf

Gump, S. (2007). "Student Evaluations of Teaching Effectiveness and the Leniency Hypothesis: A Literature Review." *Educational Research Quarterly* 30: 56-69.

Johnson, V.E. (2002). "Teacher Course Evaluations and Student Grades: An Academic Tango" *CHANCE: The Journal of the American Statistical Association,* 15: 9-16.

Kogan, L. R., Schoenfeld-Tacher, R., and Hellyer, P. (2010). "Student Evaluations of Teaching: Perceptions of Faculty Based on Gender, Position, and Rank." *Teaching in Higher Education,* 15: 623-36.

Kuzmanovic, M., Savic, G., Popovic, M., & Martic, M. (2013). "A New Approach to Evaluation of University Teaching Considering Heterogeneity of Students' Preferences." *Higher Education: The International Journal of Higher Education and Educational Planning* 66: 153-71.

MacNell, L., Driscoll, A. & Hunt, A.N. (2015) "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching" *Innovative Higher Education* 40: 291-303. DOI: https://doi.org/10.1007/s10755-014-9313-4).

McNatt, D. B. (2010). "Negative reputation and biased student evaluations of teaching: Longitudinal results from a naturally occurring experiment." *The Academy of Management Learning and Education* 9: 225-242.

Mengel, F., Sauermann, J. & Zölitz, U. (2017). "Gender Bias in Teaching Evaluations" *Journal of the European Economic Association*.

Mitchell, Kristina M.W. & Martin, Jonathan, (2018). "Gender bias in student evaluations." *The Teacher section in PS: Political Science & Politics*. DOI: https://doi.org/10.1017/S104909651800001X

Mitchell, K. (2018). " Student Evaluations Can't Be Used to Assess Professors" Slate.com March 19, 2018 https://slate.com/human-interest/2018/03/student-evaluations-are-discriminatory-againstfemale-professors.html

Naftulin, D., Ware, J. E., Jr., and Donnelly, F. A. (1973). "The Doctor Fox Lecture: A Paradigm of Educational Seduction" *Journal of Medical Education*, 48: 630-635.

Newton, J. (2007). "What is quality? Embedding quality culture in higher education." EUA Case Studies, 17-24.

Pounder, J. S. (2007). "Is student evaluation of teaching worthwhile?: An analytical framework for answering the question." *Quality Assurance in Education* 15:178-191. DOI: http://dx.doi.org/10.1108/09684880710748938

Rantanen, P. (2013). "The Number of Feedbacks Needed for Reliable Evaluation. A Multilevel Analysis of the Reliability, Stability, and Generalisability of Students' Evaluation of Teaching." *Assessment & Evaluation in Higher Education* 38: 224-39.

Ray, B., Babb, J. and Adams Wooten, C. (2018). "Rethinking SETs: Retuning Student Evaluations of Teaching for Student Agency" *Composition Studies* 46: 34–56.

Reid, L. (2010). "The Role of Perceived Race and Gender in the Evaluation of College Teaching on RateMyProfessors.com" *Journal of Diversity in Higher Education* 3: 137–152.

Reisenwitz, T. H., (2016). "Student Evaluation of Teaching." *Journal of Marketing Education*, 38: 7-17.

Schuck, S., Gordon, S., and Buchanan, J. (2008). "What are we missing here? Problematising wisdoms on teaching quality and professionalism in higher education." *Teaching in Higher Education* 13: 537-547. DOI: http://dx.doi.org/10.1080/13562510802334772

Spooren, P., Brockx, B. & Mortelmans, D. (2013). "On the Validity of Student Evaluation of Teaching: The State of the Art" *Review of Educational Research December* 83: 598–642. DOI: 10.3102/0034654313496870

Tran, N. D. (2015). "Reconceptualisation of approaches to teaching evaluation in higher education." *Issues in Educational Research* 25: 50-61. http://www.iier.org.au/iier25/tran.html

Uttl, B., White, C. A., and Gonzalez, D. (2017). "Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related" *Studies in Educational Evaluation* 54: 22-42.

Youmans, R. J., and Jee, B. (2007). "Fudging the Numbers: Distributing Chocolate Influences Student Evaluations of an Undergraduate Course." *Teaching of Psychology* 34: 245-47.

Zabaleta, F. (2007). "The use and misuse of student evaluations of teaching." *Teaching in Higher Education* 12: 55-76. DOI: http://dx.doi.org/10.1080/13562510601102131