Professional Standards Committee (PSC) 2017-2018 Year-End Report

Committee Members: David Andresen, Kris Bartanen, Denise Despres, Fred Hamel, Suzanne Holland, Andreas Madlung, Amanda Mifflin (Chair), and Jennifer Neighbors

This report is divided into four parts: 1) work completed by the PSC in response to official charges by the Faculty Senate; 2) work on standing charges, 3) additional work in response to requests from departments and individuals, and 4) ideas for future charges.

PART 1: SENATE CHARGES

Charge 1: Review the "Faculty Opportunity Hire Policy" initiated and endorsed for a two-year term in 2015. Decide to endorse, endorse for a set term, or not to endorse the policy.

Report: The PSC unanimously renewed their endorsement of the Faculty Opportunity Hire Policy for five more years beginning with the 2017-18 academic year.

Charge 2: Self-designate charges as the PSC sees fit according to the last year-end report.

Report: The 2016-17 PSC identified the following as potential items to address in 2017-18:

1. Continuing work on bias in the student evaluation process and reassessing the student evaluation process as a whole
2. Exploring whether the labs in the sciences should be evaluated separately from the courses with which they are associated
3. Reviewing departmental evaluation criteria from History, Economics, Classics, German Studies, Geology, and Hispanic Studies

Item 1 was directly related to this year's Senate Charge 3, and was addressed therein. Item 2 was not explicitly discussed given the broader discussion of how student evaluations will be used as a whole going forward. Review of departmental evaluation criteria is discussed below in the "Standing Charges" section.

Charge 3: Continue to address the issue of bias in the student evaluation process, and recommend one or more options for addressing bias on an interim or long-term basis. Share your findings with the Committee on Diversity so that that committee can draft introductory language for the administration of evaluations.

Report: The PSC identified this charge as its priority for the year due to the importance and urgency of the issue. The committee discussed the importance of student enfranchisement, the departmental disparities in requirements for faculty review visitations, and the need for information on alternative evaluation methods currently in use at other institutions (particularly our NW5 peers). Two members of the PSC conducted extensive literature searches on the issue of bias in student evaluations of teaching (SETs), and summaries are attached as Appendix A. Available evidence suggests they do not necessarily reflect teaching skill, and are biased against

women and minority populations. The currently used evaluations were described as "damaging and discriminatory" towards faculty, especially faculty of color and women.

The PSC met with the Committee on Diversity (CoD), including Title IX officer Michael Benitez, at its March 5 meeting to discuss the issues regarding bias in SETs, what the best practices are, and what action steps might be taken. Most CoD members felt the forms cannot be fixed with re-wording, and that the key issues are how evaluations get used, and the design of the instrument itself. PSC members expressed general consensus that it is not just how the university uses SETs in the evaluation process, but what it means for the person who has to read them and use them (with no protections for the faculty member). PSC members believe that the SETs have a deep ethical bias and that they undercut every other initiative we are trying to make at Puget Sound in terms of diversity affirmation and diversity faculty hires. Some of the main concerns identified by the PSC regarding the issue of SETs in evaluation are outlined below:

- concern about the loss of enfranchisement of students if SETs are eliminated
- general agreement that the university needs to do more to provide institutional support to junior faculty and especially women and people of color, as well as to educate all faculty on the difficulties faced by these faculty regarding inherent biases in SETs
- a process for informed peer review of faculty teaching
- discipline-specific evaluations to acknowledge pedagogical differences
- an SET on the model of Berkeley's that provides a selective menu of questions and includes a section on student self-assessment
- a change in the weight of the current forms in the evaluation process
- an evaluation process that acknowledges the diversity of faculty teaching loads in the Core and lower-level classes
- The need for an efficient method of student evaluation
- The need for a mentoring system for junior and diversity faculty that separates teaching evaluation from the advancement process
- replacing student-written faculty evaluations with another method of assessing teaching

The PSC felt that the task of revamping course evaluations would require a large effort involving faculty from many disciplines as well as students. To this end, the PSC brought the issue of bias in SETs to the full faculty, in collaboration with the Faculty Senate, at the April 4 faculty meeting. David Andresen presented data regarding effectiveness and bias in SET, as well as potential actions that could be taken to mitigate the problem. The PowerPoint presentation from the faculty meeting is attached as Appendix B. The PSC asked the faculty to consider the issues and potential solutions in preparation for further discussion at the April 25 faculty meeting. The PSC presented the following rationale and motion at the April 25 faculty meeting:

> Rationale: This year, the Senate charged the PSC with addressing the issue of bias in the student evaluation process, and to recommend one or more options for addressing bias on an interim or long-term basis and to reassess the student evaluation process as a whole. The PSC has spent much of its time this year working on this charge, gathering literature on bias in student evaluations, seeking out information on how peer institutions are addressing the issue, and

consulting with the Committee on Diversity. At the April 4 faculty meeting, the PSC presented their work on this issue and concluded that bias in student evaluations of teaching is a problem. The issue of implementing a long-term solution warrants further work, and is beyond the scope of the PSC. For this reason, the PSC recommends that the Senate form an ad hoc committee for the purpose of identifying a long-term solution to the problem of bias in student evaluations.

Motion: The members of the 2017-18 Professional Standards Committee move that the Faculty Senate create an *ad hoc* committee for the purposes of

1) mitigating the problem of bias in student evaluations, and

2) recommending a long-term solution or change to our current system.

The motion passed unanimously. As the Senate begins its work to create the *ad hoc* committee, the PSC would like to provide a brief summary of its discussions on the potential make-up and charges for the committee for their consideration. The committee might include two members from: the Faculty Senate, the PSC, the FAC (former members), the Committee on Diversity, student body, and at large. The PSC debated at length the pros and cons of student membership on this committee and alternative forms of student involvement. Some PSC members felt strongly that students should not be on the *ad hoc* committee (at least in the initial stages) due to concern that faculty may not feel comfortable being completely honest in the presence of students. Other members thought that transparency with students was important and they should be on the committee. The PSC did feel strongly that if students are not on the *ad hoc* committee, they should be consulted at regular intervals, perhaps in the form of focus groups. PSC and faculty meeting discussions presented a number of potential actionable interim measures, listed below:

- the education of department chairpersons and all head officers of reviews on the nature of SET Bias through formal conversations with PSC members
- the addition of language to the current SET acknowledging the problem of bias
- the creation of a mentorship program to guide junior faculty through the evaluation process (mentors outside the home department might be selected and assigned by the Provost); mentors, in turn, would provide a letter for the reviewee's file contextualizing the SET evaluations after working through them with the mentee.
- the addition of language acknowledging SET Bias to the University Faculty Evaluation Criteria and Procedures document
- the implementation of mid-term course evaluations to provide faculty (including visiting faculty) with actionable information for course improvement.
- Members of the PSC meet with each department to educate faculty members on SET bias and suggest how the form (should we continue to use the current one) should be read and used in the review process.

The PSC anticipates including language in the User Guide regarding the issue of bias in SETs as an interim measure to mitigate the problem.

Charge 4: Reassess the student evaluation process as a whole.

Report: This charge was addressed through the PSC work on Charge 3.

PART 2: STANDING CHARGES

Review Cycle for Departmental Evaluation Standards: Last year, the PSC established a review cycle whereby each university department will be asked to review and revise its departmental evaluation procedures (i.e. its departmental guidelines for promotion, tenure, and other reviews) every eight years. The departments scheduled for review for 2017-18 were Classics, German Studies, Geology, and Hispanic Studies.

- The PSC reviewed departmental evaluation standards for Economics, Hispanic Studies, History, and Occupational Therapy (Economics and History were carried over from last spring, and OT submitted changes voluntarily). Changes and suggestions were discussed and will be forwarded to the respective departments. Each department will be asked to return revisions to their evaluation document in Fall 2018.

- Math and Computer Science submitted a revised departmental evaluation standard document, and it was approved by the PSC on May 9.

- The PSC has not received the departmental evaluation standards from Classics, German Studies, and Geology. Classics and German Studies requested extensions, which were granted.

Interpretations of the Faculty Code:

Interpretation of Faculty Code Chapter III, Section 4, f (1)

The PSC met on Friday, February 16, 2018, regarding a requested interpretation of Faculty Code Chapter III, Section 4, f (1), and specifically whether a faculty evaluation that has gone to the presidential level can be suspended in response to a concern/grievance regarding the professional ethics of an evaluator at the departmental or Faculty Advancement Committee levels. The interpretation was found to be significant and was submitted to the Faculty Senate Chair on March 15, 2018 and is attached to the minutes of the March 19, 2018 Faculty Senate meeting.

The PSC finds that an evaluation process at the presidential level can indeed be suspended for a concern/grievance regarding professional ethics. The PSC, in its reading of Faculty Code Chapter III, Section 4, f (1), finds that the evaluation process is not limited to the levels noted in lines 34 and 35 of the text ("the department, program, school, or Faculty Advancement Committee level"), and is ongoing until the Board of Trustees has rendered a final decision.

The PSC also finds this interpretation to be significant, because it bears upon the definition of “evaluation process” in the Code, especially as this relates to different kinds of grievances. The PSC notes that this interpretation applies just to grievances of professional ethical behavior, as

referred to in Chapter III, Section 4, f (1), and does not supersede the process the code provides for other sorts of appeal.

The PSC plans to insert the following language in the appendix to the Faculty Code:

Interpretation of Chapter III, Section 4. Evaluation procedure

1. The evaluation process is considered ongoing until the Board of Trustees has rendered a final decision.

Interpretation of Chapter III, Section 4, f (1). Process for dealing with questions of professional ethics that arise during an evaluation

1. The evaluation process can be suspended at any time until the evaluation is complete, including at the level of the President and Board of Trustees, when due to a grievance arising from concerns about professional ethical behavior of an evaluator at the departmental or Faculty Advancement Committee levels.
2. This interpretation applies just to grievances of professional ethical behavior, as referred to in Chapter III, Section 4, f (1), and does not supersede the process the code provides for other sorts of appeal.

PART 3: ADDITIONAL WORK

- Robin Jacobson and Seth Weinberger visited the October 20 PSC meeting to discuss the status of the proposal for a BA degree for Washington Corrections Center for Women students. They gave a brief overview of the FEPPS program as it currently exists and described a proposal process for a Puget Sound-sponsored BA degree for the future. Robin and Seth did not seek a formal endorsement from the PSC, but wanted to hear what the committee thought about possible points of discussion that might come up about issues that might fall under the purview of the PSC in such future program.

- The PSC was asked if they could help clarify the specific description of the position that is currently held by Dean (of the University) Bartanen and Dean (of the Faculty) Kukreja. The inquiry was made in the context of transparency, since there has been some shuffling of responsibilities of the Dean of Students, Dean of the University, and Dean of the Faculty. Given the temporary nature of Dean Bartanen's role as Dean of Students, the PSC decided that no action on this issue was necessary at this time.

- The committee discussed a Faculty Senate proposal (see Appendix C) to revise the language of the Faculty Code regarding promotion to full professor. The committee discussed the proposed language and offered suggestions for the Senate to consider.

- The committee is in the process of discussing revisions to the *Faculty Evaluation Procedures and Criteria* document (User Guide). Revisions will be completed by July 1.

PART 4: FUTURE CHARGES

The work that the PSC hopes to address in the 2018-19 academic year includes:

- Review of departmental evaluation criteria according to the published review cycle. Evaluation standards from Classics, German Studies, and Geology remain outstanding from previous review cycles. Evaluation standards from Religious Studies, Exercise Science, Psychology, and Sociology and Anthropology are scheduled for review in 2018-19.

- Addressing a request from a faculty member in the School of Education regarding the clinical streamlined instructor review process that the 2017-18 PSC did not have time to discuss. The PSC was asked to consider the language on page 27 of the *Faculty Evaluation Procedures and Criteria* document that addresses the streamlined instructor review process. The request from the faculty member is summarized below:

  The relevant line is, "Instructors who have served 17 years or more in that rank may establish an alternating schedule of full and alternative reviews in consultation with the head officer and the Dean under the procedures described in this section." Since tenure-line professors typically go up for a full review at year 11 (about a decade), the faculty member wonders if after instructors pass that timeline (10 years of service), they could then be eligible for alternating reviews. Since instructors are up for review <u>every three years</u>, (compared to 5 for professors), there is already a significant check about their teaching. Perhaps after a decade the instructor review could be every 5 years, instead of every three. The faculty member feels that the current cycle of review seems like a lot to ask of long term clinical faculty.

- Addressing the evaluation process for non-tenure-line positions, including visiting faculty members that stay beyond 3 years and other potential renewable non-tenure-line positions that may be created. Former Associate Dean Martin Jackson initiated a conversation about this issue with the PSC, FSC, and Senate in 2016, but has not been able to follow up due to time constraints. The issue is currently impacting a number of departments on campus, so we recommend that the Senate reach out to Martin Jackson for additional context and ask the appropriate committees to address the problem.

Respectfully submitted on behalf of the PSC,

Amanda Mifflin, Chair

Appendix A

Summary of Literature Review on Bias in SETs by Denise Despres

In May 2017, reference librarian Andrea Kueter provided me with a selective list of recently published materials on bias in student evaluations of teaching. Below you will find the citations and abstracts for scholarly articles that might serve as a preliminary basis for the PSC discussion of whether student evaluations of teaching are appropriate material to include in personnel files for hiring, tenuring, and advancing staff and faculty. All of these articles acknowledge that this is a new field of inquiry for research in the fields of Education, Psychology, and other disciplines.

1) Smith, Bettye P; Johnson-Bailey, Juanita. "Student Ratings of Teaching Effectiveness: Implications for Non-White Women in the Academy." Negro Educational Review. Vol. 62/63 (2011-1012): 115-140, 266.
The purpose of this study was to describe student ratings of teaching effectiveness for women faculty at a Southern Research Extensive University. Of the 82 women faculty in this study, 61 or 74% were white, 13 or 16% were Black, and 8 or 10% were identified as "Other" (including Asians, Latinos, and Native Americans). Both undergraduate and graduate level courses were used to analyze student ratings for 28 items used for the end-of-course evaluation. Of these 28 items, 26 were multidimensional and addressed specific topics or a single aspect about instruction and 2 were global, which addressed the overall value of the course and overall teaching ability. The finding showed that non-White female faculty in this study had above average mean scores on the multidimensional and global items. White female faculty had higher mean scores that female faculty identified as "Other" and Black female faculty on all items, multidimensional and global. Also, there was a significant difference between the mean scores of White female faculty and Black female faculty on multidimensional and global items (publication abstract)
http://ezproxy.ups.edu/login?url=http://search.proquest.com/docview/940916203?accountid=1627

2) Smith, Bettye P; Hawkins, Billy. "Examining Student Evaluations of Black College Faculty: Does Race Matter?" The Journal of Negro Education, Vol 80.2. (2011): 149-162.
The purpose of this study was twofold. First, to describe the undergraduate student ratings of teaching effectiveness based on the traditional 36-item end-of-course evaluations form used in the College of Education (COE) at a southeastern Research Extensive predominantly White institution. Second, using critical race theory (CRT) to compare the teaching effectiveness for the tenure track faculty in this study based on race (White, Black and Other racial groups including Asians, Latinos, and Native Americans). Three academic years of undergraduate level courses were used to analyze student ratings for 28 items (26 multidimensional, which address specific topics or a single topic about instruction and 2 global/overall, which address the value of the course and teaching ability) on the end-of-course evaluations form. Eight of the 36 items request demographic information from the student. The findings showed that of the three racial groups, Black faculty mean scores were the lowest on the 26 multidimensional items. On the two global items, which are used in making personnel decisions, Black faculty m mean scores were also the lowest of the faculty groups analyzed. (publication abstract)
http://ezproxy.ups.edu/login?url=http://search.proquest.com/docview/909492695?accountid=1627

3) MacNeill, Lillian; Driscoll, Adam; Hunt, Andrea N. "What's In a Name: Exposing Gender Bias in Student Ratings of Teaching." Innovative Higher Education. Vol. 40.4 (2015): 291-303.
Student ratings of teaching play a significant role in career outcomes for higher education instructors. Although instructor gender has been shown to play an important role in influencing student ratings, the extent and nature of that role remains contested. While difficult to separate gender from teaching practices in person, it is possible to disguise an instructor's gender identity online. In our experiment, assistant instructors in an online class each operated under two different gender identities. Students rated the male identity significantly higher than the female identity, regardless of the instructor's actual gender, demonstrating gender bias. Given the vital role that student ratings play in academic career trajectories, this finding warrants considerable attention. (publication abstract)
http://ezproxy.ups.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1068677&site=ehost-live&scope=site

4) Schueths, April M.; Gladney, Tanya; Crawford, Devan M.; Bass, Katherine L.; and Moore, Helen A. "Passionate pedagogy and Emotional labor: Students' responses to learning diversity from diverse instructors." International Journal of Qualitative Studies in Education. Vol. 26.10 (2013): 1259-1276.
This qualitative study examines emotion themes reflected in student evaluations from required diversity courses at a predominantly white, US public university. We analyze two years of student evaluations for 20 instructors. Situated by the work of Acker, Jaggar, and Hochschild, we find contradictory themes of perceived instructional bias and the value of diversity lessons. Student evaluations result in systematic disadvantage for minority instructors that may be heightened for female instructors of color. Non-minority instructors (both male and female) gain privileges by avoiding dealing with diversity directly which is reflected in student evaluations through the process of "ducking diversity". The organizational structure required of diversity courses marginalized the scholarship and emotion work of minority instructors and inherently reproduced the very inequalities they are designed to combat. (publication abstract)
http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1241&context=sociologyfacpub

5) Boring, Anne; Ottoboni Kellie; and Stark, Philip B. "Student evaluations of teaching (mostly) do not measure teaching effectiveness." Boring et al. ScienceOpen Research 2016 (DOI: 10. 14293/S2199-1006.1.SOR EDU.AETBZC.v1)
Student evaluations of teaching (SET) are widely used in academic personnel decisions as a measure of teaching effectiveness. We show:
. SET are biased against female instructors by an amount that is large and statistically significant. . The 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. . It is not possible to adjust for the bias, because it depends on so many factors. . SET are more sensitive to students' gender bias and grade expectations than they are to teaching effectiveness. . Gender biases can be large enough to cause more effective instructors to get lower SET than less effective instructors.
These findings are based on nonparametric statistical tests applied to two datasets: 23,001 SET of 379 instructors by 4,423 students in six mandatory first-year courses in a five-year natural experiment at a French university, and 43 SET for four sections of an online course in a randomized, controlled, blind experiment at a US university. (publication abstract)

https://www.math.upenn.edu/~pemantle/active-papers/Evals/stark2016.pdf

6) Basow, Susan A.; Martin, Julie L. "Bias in Student Evaluations."
Effective Evaluation for Teaching: A Guide for Faculty and Administrators. Ed. Mary E. Kite. Society for the Teaching of Psychology: 2012.
Given the important role student evaluations play in many academic employment decisions—such as hiring, promotion, tenure, salary, and awards—it is vital to understand potential sources of bias. In this chapter, we will examine potential biasing factors involving the professor—such as gender, race/ethnicity, attractiveness, and age—as well as the course, such as course difficulty and expected grade . . . Although there is still a considerable amount of research needed to understand all the ways that student evaluations can be biased, this chapter suggests that not only is some bias possible but it is likely. As a human activity reliant upon person perception and interpersonal judgment, student ratings are affected by the same factors that can potential affect any raters judgment: stereotypes based on gender, race/ethnicity, age, and other qualities (such as professor sexual orientations); the equations of "what is beautiful is good;" more positive feelings towards those who seem to reward us (e.g., with good grades). Even though the size of individual effects may be small, for specific professors these small effects may add up to make a meaningful difference on the ratings they receive. Although the average-looking young-to-middle-aged White male professor teaching traditional courses may receive student ratings that are relatively unbiased reflections of his teaching effectiveness, other professors (women, minorities, older, unattractive-looking, teaching diversity-related courses) may receive evaluations that reflect some degree of bias. It behooves those who use such ratings for evaluative purposes to understand the subtle ways such variables may operate, especially in interaction with each other. (excerpt and summary)
https://dspace.lafayette.edu/bitstream/handle/10385/1405/Basow-EffectiveEvaluationofTeaching-2013.pdf?sequence=1

7) Vaillancourt, Tracy. "Students Aggress Against Professors in Reaction to Receiving Poor Grades: An Effect Moderated by Student Narcissism and Self-Esteem." Aggressive Behavior. Vol 39.1 (2013): 71-84.
Laboratory evidence about whether students' evaluations of teaching (SETs) are valid is lacking. Results from three (3) independent studies strongly confirm that "professors" who were generous with their grades were rewarded for their favor with higher SETs, while professors who were frugal were punished with lower SETs (Study 1, d = 1.51; Study 2, d = 1.59; Study 3, partial $\eta 2$ = .26). This result was found even when the feedback was manipulated to be more or less insulting (Study 3). Consistent with laboratory findings on direct aggression, results also indicated that, when participants were given a poorer feedback, higher self-esteem (Study 1 and Study 2) and higher narcissism (Study 1) were associated with them giving lower (more aggressive) evaluations of the "professor." Moreover, consistent with findings on self-serving biases, participants higher in self-esteem who were in the positive grade/feedback condition exhibited a self-enhancing bias by giving their "professor" higher evaluations (Study 1 and Study 2). The aforementioned relationships were not moderated by the professor's sex or rank (teaching assistant vs. professor). Results provide evidence that (1) students do aggress against professors through poor teaching evaluations, (2) threatened egotism among individuals with high self-esteem is associated with more aggression, especially when coupled with high narcissism, and (3) self-enhancing biases are robust among those with high self-esteem. Aggr. Behav. 39:71-84, 2013. © 2012 Wiley Periodicals, Inc. [ABSTRACT FROM AUTHOR]

http://ezproxy.ups.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=85017042&site=ehost-live&scope=site

8) Boysen, Guy A.; Kelly, Timothy J., Raesely, Holly N.; Casner, Robert W. "The (mis)representation of teaching evaluations by college faculty and administrators." Assessment & Evaluation in Higher Education. Vol. 39.6 (2014): 641-656.
Student evaluations of teaching are ubiquitous and impactful on the careers of college teachers. However, there is limited empirical research documenting the accuracy of people's efforts in interpreting teaching evaluations. The current research consisted of three studies documenting the effect of small mean differences in teaching evaluations on judgements about teachers. Differences in means small enough to be within the margin of error significantly impacted faculty members' assignment of merit-based rewards (Study 1), department heads' evaluation of teaching techniques (Study 2) and faculty members' evaluation of specific teaching skills (Study 3). The results suggest that faculty and administrators do not apply appropriate statistical principles when evaluating teaching evaluations and instead use a general heuristic that higher evaluations are better. [ABSTRACT FROM PUBLISHER]
http://ezproxy.ups.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=96539445&site=ehost-live&scope=site

9) Clayson, Dennis E.; Haley, Debra A. "Are Students Telling Us the Truth? A Critical Look at the Student Evaluation of Teaching." Marketing Education Review. Vol. 21.2. (2011): 101-112.
Over 99 percent of business schools use student evaluation of instruction to measure teaching and classroom performance. The resultant measures are utilized in judgments of merit pay, tenure, and promotion. In such an environment, an inspection of exceptions to their assumed validity is justified. This paper investigates one such issue that is rarely reported. Simply put: to what extent are students telling us the truth when they evaluate instruction? A literature review indicates that students (1) ignore or falsify answers in light of variables considered more important, (2) give subjective impressions in response to objective questions, and (3) at times give purposefully misleading and false responses. A survey of students found that a majority knew of respondents who had falsified evaluations and that an estimated 30 percent of evaluations contain answers the students knew were not true. In light of these findings, the validity of student evaluation of teaching to improve individual instructor performance, modify curriculum, and create comparative scales to evaluate faculty is called into question. (publication abstract)
http://ezproxy.ups.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ972749&site=ehost-live&scope=site

Summary of Literature Review on Bias in SETs by David Andresen:

Background: The PSC was concerned about use of student evaluations of teaching (SETs) given studies showing gender and race bias in SETs. However, even if bias issues *could* be addressed, a more fundamental issue remains of whether or not SETs are valid measures of teaching effectiveness in the first place. The following is my (David Andresen) assessment regarding use of SETs for evaluation purposes.

Q: Do SETs tell us about teaching effectiveness?

Student evaluations of teaching (SET) being used as an indicator of teaching effectiveness was thoroughly review in a recent paper in *Studies in Educational Evaluation* (Uttl, White, & Gonzalez, 2016) and reanalyzed using modern meta-analytic methods. Here, teaching effectiveness means student performance on the final exam. After examining 51 articles and with 97 multisection studies (i.e., students randomly assigned to different instructor sections of the same course), this study produced two clear results:

1) Prior foundations studies that support a link between SET and student learning (Cohen, 1981; Clayson, 2009; Feldman, 1989) were flawed in several fundamental ways (e.g., artifact effects of small sample sizes, inappropriate analyses, publication bias, etc.) and after appropriate re-analyses they do not, in fact, support a link between SET and student performance
2) The metaanalysis of all work in this area examining the relation between SET and final exam performance show no relation.

In another study (Bark, 2014), researchers looked not at performance in the current class, but how students did in the *next* class in the sequence. In other words, do SETs from a 101 course predict success of that student in the 102 course? The results showed that there was a *negative* correlation between SET and success in the next course in the sequence. In other words, the teachers who got *lower* SET ratings produced students who did *better* in subsequent courses. In this case, high SET ratings actually predicted that students would do *worse* later on.

Also keep in mind that other literature reviews in this area that appear to support SET as a measure of learning are based largely on the work of Cohen (1981) and Feldman (1989), which now have been shown to be incorrect in the first place.

Conclusion: The use of SETs to evaluate teaching *with regard to measureable learning outcomes* is not supported by evidence.

Q: What do other universities do for teaching evaluation?

Although some universities do not use SETs (e.g., Reed appears to use solicited student letters), most universities still use them. However, most universities also emphasize that SETs should only be part of the evaluation of teaching in conjunction with other approaches (e.g., course visits,

self-evaluation). The evidence suggests the numerical ratings from SETs should be given very little weight.

Q: If SETs don't tell us about teaching, what do they tell us about?

SETs are not really teaching evaluations. Instead, they are student *perception* surveys. If we keep in mind that SETs tell us how a student *feels* about a course given both relevant (e.g., teaching) and irrelevant (e.g., gender of teacher), we can begin to place the appropriate weight to SETs.

The numerical ratings are the main concern of SETS, written comments have not been well-studied but are certainly more informative if not much more valid. For example, a student who gives every item on a SET 1's does not really provide any information about why they did that. However, a written comment forces the student to justify their perceptions and can be helpful.

Q: So how do we best assess teaching effectiveness?

No easy answer for this. However, we are all teachers and we should probably incorporate many more peer class visits both before and during evaluations, making regular visits to colleagues' classes part of our culture. Writing a promotion letter based on one or two class visits means much of the actual content comes from the SETs, which we know aren't great, and the impact of the numerical ratings should be minimized.

Q: Students being part of the process is important. How do we include student views on teaching?

The main issue with SETs is about trying to assess anything about teaching effectiveness from numerical ratings. If we want to keep numerical ratings, then some things that would make them better are:

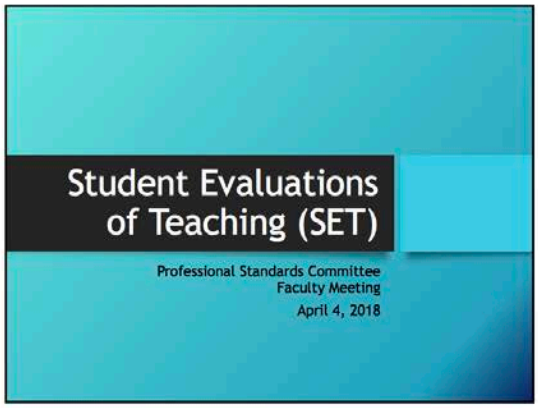
1) Making sure averages of a professor's numerical ratings should be accompanied by indicators of distribution (e.g., standard deviation, range). In this way, evaluators can know if a rating of "3" means they actually got mostly 3s, or they got 1's and 5's…both would have about the same average but for very different reasons.
2) Some institutions account for student prior background, overall academic performance, etc. in a complex statistical model to account for variability of this type and reveal more accurate ratings. This could be done, but someone would have to do it and probably would cost money. I couldn't find a lot about how much better these ratings relate to teaching effectiveness, so difficulty to know how much better they really are.

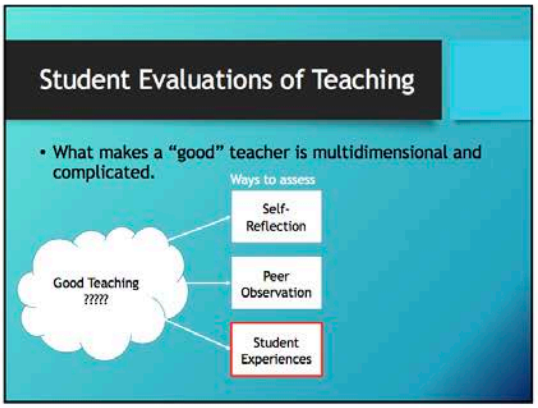
If we wanted to scrap the numerical scale but keep SETs, we could:

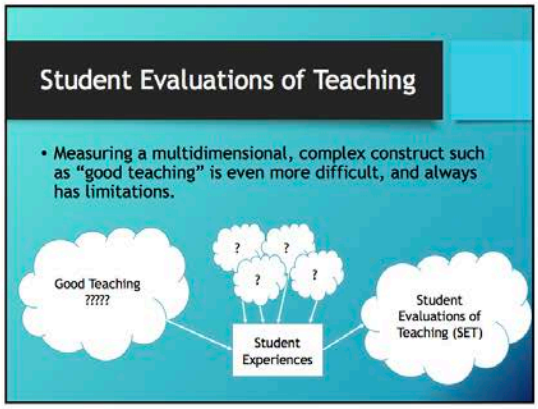
1) Base everything on analysis of written comments of students to see patterns in the comments. We may need to develop a new SET with questions that directly solicit information we want to know.
2) Another approach in the literature is to use a best-worst ranking of various aspects of teaching. So no numerical ratings for individual items, but students take a list of terms

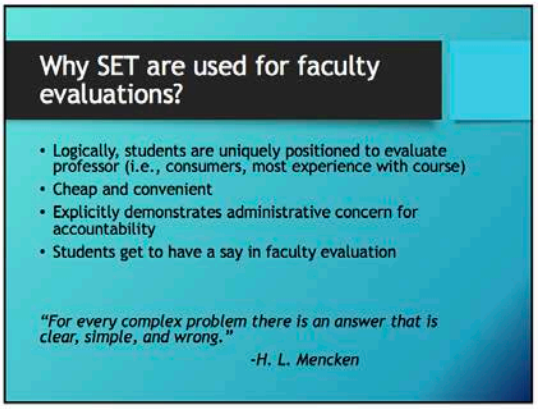
and phrases related to teaching (e.g., material was interesting, expectations were clear, etc.) and put them in order from best to worst. By doing this, students are forced to say what was both good and bad about the course regardless of how they feel about it overall; there's no way to "give all 5s" or "give all 1's" which aren't very helpful.

Appendix B


Student Evaluations of Teaching (SET)
Professional Standards Committee
Faculty Meeting
April 4, 2018


Student Evaluations of Teaching
• What makes a "good" teacher is multidimensional and complicated.
Ways to assess
Good Teaching
?????
Self-Reflection
Peer Observation
Student Experiences


Student Evaluations of Teaching
• Measuring a multidimensional, complex construct such as "good teaching" is even more difficult, and always has limitations.
Good Teaching
?????
Student Experiences
Student Evaluations of Teaching (SET)


Why SET are used for faculty evaluations?
• Logically, students are uniquely positioned to evaluate professor (i.e., consumers, most experience with course)
• Cheap and convenient
• Explicitly demonstrates administrative concern for accountability
• Students get to have a say in faculty evaluation
"For every complex problem there is an answer that is clear, simple, and wrong."
-H. L. Mencken

Problems Using SET

- SET ratings *do not* relate to teaching effectiveness (as measured by exam performance)
- SET ratings *do reflect* gender and potential racial bias
- Unsettling given the importance of SET in promotion and tenure

Prior evidence that SET do relate to teaching effectiveness

- Uttl, White, & Gonzalez (2016)
  - Meta-analysis of 51 studies of SET
  - Essentially an analytical review of the literature
- Two important conclusions
  1. Prior results that are *repeatedly* cited as evidence supporting SET (e.g., Cohen, 1981; Feldman, 1989) had *serious* methodological issues
  2. Scientifically-sound studies show no correlation between SET and outcomes

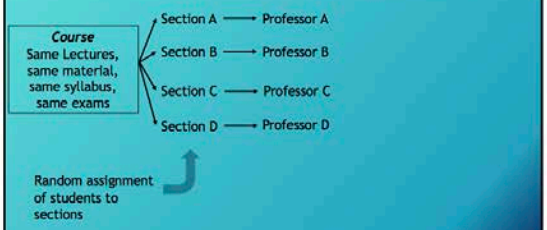
Best *Scientific* Evidence

- To best assess teaching skill, we need an objective, empirical measure of teaching effectiveness: exam scores
  - Assumptions:
    - On average, exam scores reflect learning of course material
    - On average, learning of course material is related to teaching effectiveness
    - Thus, on average, exam scores should reflect teaching effectiveness
- Because course topic, level, and focus all impact exam scores, the best studies use multiple sections of the same class

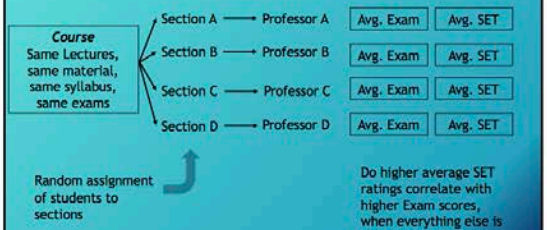
Best Experimental Evidence

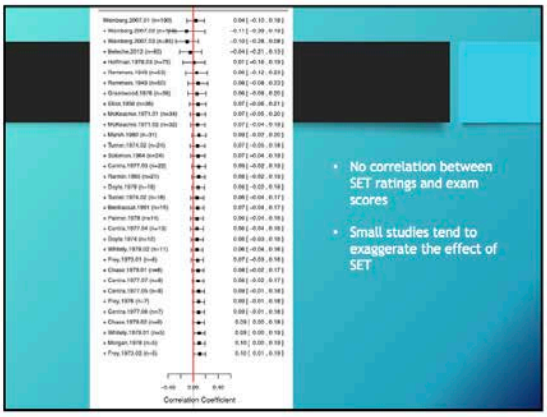
Course
Same Lectures, same material, same syllabus, same exams

Best Experimental Evidence



Best Experimental Evidence





SET and student learning

- Perhaps SET could be useful for other purposes, but they do not tell us about teaching effectiveness (as defined by objective exam performance)
- So what *do* they measure? Evidence suggests more of a subjective "feeling of satisfaction" measure, rather than teaching effectiveness
  - Affected by many other non-teaching factors, including unrelated, subjective characteristics such as attractiveness (Hammermesh & Parker, 2005)

SET and bias

- Given that SET reflect feelings, not surprisingly, both explicit and implicit biases are reflected in SET

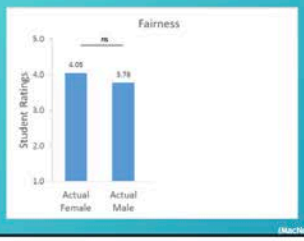
SET and gender bias

- Clearest evidence of gender bias comes from online courses in which *the perceived gender* of the instructor was manipulated
  - Male instructor
    - Taught sections as a professor with a male name
    - Taught sections as a professor with a female name
  - Female instructor
    - Taught sections as a professor with a male name
    - Taught sections as a professor with a female name
  - Allows analysis of SET ratings for actual male and female instructor, as well as *perceived* male and female instructor

(MacNell, Driscoll, & Hunt, 2014)

SET and gender bias

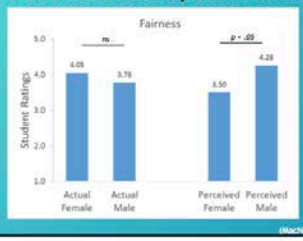
- No difference between the *actual* female and male ratings



(MacNell, Driscoll, & Hunt, 2014)

SET and gender bias

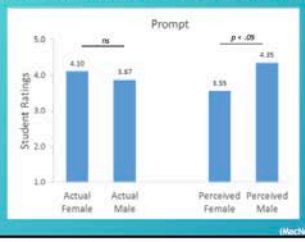
- No difference between the *actual* female and male ratings
- Significant difference between *perceived* female and male



(MacNell, Driscoll, & Hunt, 2014)

SET and gender bias

- Significant difference between *perceived* female and male even for relatively objective aspects of teaching



(Mechell, Driscoll, & Hunt, 2014)

Table 4
Unpaired T-Test of SET by Category

	N	Mean Rating	Difference	T	P
Instructor					
Dr. Martin (man)	255	3.84	0.4***	5.24	0.000
Dr. Mitchell (woman)	835	3.44			
Instructor/Course					
Dr. Martin	255	3.71	0.4***	4.63	0.000
Dr. Mitchell	835	3.31			
Course					
Dr. Martin	357	3.71	0.22***	3.11	0.001
Dr. Mitchell	1,169	3.49			
Technology					
Dr. Martin	153	3.83	0.19**	1.93	0.027
Dr. Mitchell	501	3.64			

(Mitchell & Martin, 2018)

Table 4
Unpaired T-Test of SET by Category

	N	Mean Rating	Difference	T	P
Instructor					
Dr. Martin (man)	255	3.84	0.4***	5.24	0.000
Dr. Mitchell (woman)	835	3.44			
Instructor/Course					
Dr. Martin	255	3.71	0.4***	4.63	0.000
Dr. Mitchell	835	3.31			
Course					
Dr. Martin	357	3.71	0.22***	3.11	0.001
Dr. Mitchell	1,169	3.49			
Technology					
Dr. Martin	153	3.83	0.19**	1.93	0.027
Dr. Mitchell	501	3.64			

Table 3
Grading Averages

	Dr. Martin (Man) Course Averages	Dr. Mitchell (Woman) Course Averages
Final Grades	75.23	79.30
Discussion Posts	67.97	73.09
Short Answers	66.60	67.74

(Mitchell & Martin, 2018)

SET and race bias

- Very few studies, but follows from gender bias evidence that other biases would also impact SET
- Study of 25 highest-ranked liberal arts colleges faculty on RateMyProfessor

Table 2
Instructor Ratings for Racial Minority and White Faculty

Rating	Racial minority	White	$F(1, 3550)$
Overall Quality	3.72 (.94)	3.89 (.87)	17.64**
Helpfulness	3.81 (.98)	3.95 (.90)	11.03**
Clarity	3.64 (.99)	3.83 (.92)	20.89**
Easiness	3.03 (.81)	2.94 (.77)	7.11**

Note. *SD*s in parentheses.
* $p < .05$. ** $p < .01$.

(Reid, 2010)

Conclusions

- SET ratings do not relate to objective measures of teaching effectiveness
- SET ratings reflect gender and potential race bias
  - Is using SET to make tenure and promotion decisions is ethical?
  - Does use of SET ratings work against University vision to "increase the diversity of all parts of our university community through commitment to diversity in our recruitment ***and retention*** efforts?"
- Using methods known to be biased to make tenure and promotion decisions is illegal, and invites lawsuits
  - Jan 2017: In lawsuits filed in federal court last week and state court last summer, an assistant professor of film and media studies accuses her employer of improperly relying in part on discriminatory feedback from her students in deeming her work subpar. (one of many examples)

A range of possible solutions

Appendix C

Request from the Faculty Senate:

The Faculty Senate has spent more than a year collecting information and developing recommendations to revise the Faculty Code with regard to the language around promotion to full professor. A committee of members of the Faculty Senate, convened by the Faculty Senate to do this work, believes it would be best to present two options to the faculty:

- a simple revision of the existing language that would clarify any ambiguity
- a more expansive revision of the language that would alter our expectations about promotion to full

Below, please find our draft language of the second of these. We are sending this to the PSC on the understanding that the PSC can offer valuable insights into the compatibility of this language with the Faculty Code (and whether other portions of the Code or elements of the review process would have to be altered as well).

PROPOSED REVISION to the Faculty Code (at III.3.e):

"Faculty promotion shall be based upon the quality of a person's performance of academic duties. Specifically, decisions whether to promote shall be based upon the quality of the faculty member's performance in the following areas, listed in order of importance: (1) teaching and related responsibilities, including the mentoring of students; (2) professional growth; (3) participation in service to the university, to one's profession, or—in ways related to one's professional interests and expertise—to the larger community. Because the university seeks the highest standards for faculty advancement, mere satisfactory performance is no guarantee of promotion. In addition, appointment in the rank of associate professor and professor normally requires a doctoral, or other equivalent terminal degree. Within the category of professional growth, candidates for promotion to the rank of full professor must demonstrate significant scholarly achievement. Within the category of service, candidates for promotion to the rank of full professor must provide evidence of a significant contribution to the university. In no case is promotion to be recommended without a determination that the candidate has maintained a consistent high quality of teaching and a sustained record of service."