TO: Faculty Senate, University of Idaho  
FROM: Teaching & Advising Committee (TeAC), Chair, Stephan Flores, Department of English  
RE: Review of Research on Possibilities for Gender Bias in Students’ Evaluation of Teaching  
DATE: 18 April 2018  

Leadership in Faculty Senate forwarded to the Teaching & Advising Committee (TeAC) separate requests from Professors Terry Grieb, College of Business and Economics (CBE) and a professor in the College of Letters, Arts, and Social Sciences (CLASS) to explore recent research regarding evidence of bias in Student Evaluation of Teaching (SET) forms. Of particular concern was bias or variance along lines of gender—that is, whether students evaluate male and female faculty members differently, including claims that female faculty are viewed more negatively on SETs than male faculty. Those requests from faculty are copied further below, at the end of this report. Also listed below, in addition to the research referenced in the requests, are multiple references to recent research on this question that members of TeAC have reviewed, in consultation with Dale Pietrzak, Director of Institutional Effectiveness & Accreditation at the University of Idaho. The committee also considered UI internal data, provided through recent analyses comparing the prior SET form to the newly implemented SET form at UI.

Also copied below are excerpts and abridged accounts of Dale Pietrzak’s analyses of the primary research studies that were referred to us, and instances of our own responses and analysis. Dale’s complete analyses and data reports are available, as are the fuller accounts of our discussions (and email exchanges), as reported in the minutes for TeAC February 28 and March 21, 2018 meetings.

Please find as follows below these sections of our report: (1) Conclusions; (2) Advancing Excellence in Assessment, Teaching, and Learning at the UI; (3) Summary-Review of Research.

(1) Conclusions

Based on review of the studies forwarded to us, several additional studies, and the recent analysis of data from UI SETs, we did not find a sufficient aggregate basis to indicate significant instances of gender bias. We found that the data, evidence, and methods provided in external studies are not in accord with the tone and degree of emphasis that popular review articles, as well as the referenced studies themselves, tend to posit and claim in their conclusions. The external studies lacked statistically substantial evidence in data sets, methodologies, and analysis to support claims of anything more than small (approximately ≤1%, even with statistical significance) variations to indicate gender bias.

This is not what we expected to find or to conclude, in part because we found the research article titles, abstracts, and conclusions frequently stressed evidence for such bias in SET forms, even if the data in the article did not support this conclusion. Some of us may expect that such biases are prevalent, especially given our sense of and scholarly expertise on
issues of gender equity and our academic interests in matters of inequality across differences of gender as well as similar and interconnected inequities across such forms of identification as race, ethnicity, and also class stratifications. However, the data do not support this belief. More specific to UI, we did not see evidence of significant variations along lines of gender in our SET forms. Based on these data, we cannot suggest that such biases are systemic and broadly structural, at least as determined by the statistical evidence across many courses and instructors.

We do expect that individual faculty may experience biases and sense that students may respond to their teaching in ways that are influenced by attitudes and values that differ according to the faculty member’s gender. In terms of what our SETs ask and ‘measure,’ however, such differences vary in numerically and statistically small ways. The literature we have reviewed also indicates small variations from bias, with related problems of sufficient data or appropriate methodology that puts such claims and results into question. For example, Ryalls et al. (cited below) described claims by Flaherty (who is citing the study by Boring et al.) as “flagrantly misleading,” that in the Boring study “gender explained no more than 1% of the variance in the ratings” and that “...correlations reported are very weak and non-significant.” Additionally, Mengel et al.’s study found some gender bias against female graduate instructors, but not for professorial faculty: “Interestingly, we do not observe this gender bias for more senior female instructors or professors” (Mengle et al. 2). As one TeAC member observed, “...most concerning is the low n....” in reference to one Inside Higher Ed-linked study with an n of just 2 instructors and a Stanford study with an n of 8. The TeAC member added that while “...there may be something worth looking into about bias, we need to be particular about what studies we use as evidence; many of them have significant methodological flaws and should not be used as support for or against a given argument.” In his comparative study of UI SETs, Dale Pietrzak concluded that in his analysis of over 3,900 courses, “Examination of the instructor and course scores with instructor gender, citizenship and minority status found no relationship to gender or minority status.”

We think that it is important to continue to explore and monitor the literature on these questions about bias, as well as periodically review UI assessment and institutional data related to teaching and its evaluation. We agree that SETs should not be the primary and/or only basis for feedback on teaching, evaluations of teaching, or reviews of teaching for tenure and promotion. As noted below, we advocate for broadening the scope of formative and summative reviews of teaching in order to advance excellence in assessment, teaching, and learning at UI. We also wish to promote testing and implementing a greater range of such methods for evaluation and review of teaching at UI. Strengthening these methods may provide better, alternative bases for considering how different biases may affect instruction and evaluation and what might be done about it.

(2) Advancing Excellence in Assessment, Teaching, and Learning at the UI

A solutions-based approach to the controversies surrounding SETs must at once acknowledge that (1) student input is important to the teaching and learning process, (2) SET data alone do not capture or directly measure instructional effort and impact, (3)
alternative forms of assessment exist, and (4) a more dynamic representation of instructional effort can (and should) be integrated into existing practices and policies. Such an approach provides an opportunity for enhanced formative/developmental support for faculty, improved administrative/summative assessments of faculty, and added value to pedagogical innovation and effort. Further, it advances a culture that at the University of Idaho, teaching and learning matter.

This effort requires attention to the current institutional ethos, a context for discussing SETs, alternative and supplemental forms of evidence-based instructional evaluation, and support systems for faculty and administrators. More specifically, this involves:

- Investigating and adopting strategies to increase response rates for SETs;
- Educating faculty and administrators on how to make the most of SETs and how to move beyond them; and
- Focusing on how we can address difference and advance change that strengthens learning across different settings and modes of delivery and engagement.

Presently, the majority of programs on campus use SET data as the single determinant of instructional impact, often nested within a narrative of teaching effectiveness. In the context of this conversation, it is necessary to identify methods to “make the most” of SETs, and to expand the range of ways that teaching is and can be supported and evaluated on campus. One way to imagine this is to envision three legs of a stool—student-evaluations, peer-evaluations, and self-evaluations—and how, together, they converge to provide evidence of instructional effort and impact. With attention to each:

I. **Student-evaluations** must be kept in perspective.

   a. Our base instrument is small—two base questions—and can be enriched through the integration of questions that are formatively useful to the instructor. These data can give the instructor something they can work with, and the Center for Excellence in Teaching and Learning (CETL/https://www.webpages.uidaho.edu/cetl/, directed by Brian Smentkowski) is here to work with faculty on dissecting responses and formulating solutions.

   b. Reliability. The reliability of an instrument is compromised when response rates are low. Fortunately, the small N is a problem that can be remedied. If faculty articulate the value of the process to students and, in the context of that conversation (or conversations), provide them with time to complete the SET form, the N will surge.

   i. Developing a shared value for the process, for the meaningfulness of student feedback, establishes buy-in and a sense that feedback matters. When this occurs, the quantity and quality of responses typically improve. In short, if faculty act like it doesn’t matter, students will (generally) act like it doesn’t matter; if faculty act like it matters, so to (generally) will their students.

   c. Validity. It is critical to remember that SETs measure student perceptions of the quality of instruction and course quality.
i. A host of factors influence SETs, and they represent a pixel in what is actually a motion picture of teaching. Even among nationally-normed instruments, the general rule of thumb is that student ratings should count no more than 30%-50% of teaching evaluation.

d. So, we—particularly faculty and administrators—can “make the most of student evaluations” within appropriate parameters, but they should not be used as the indicator of instructional effectiveness.

Because SETs do not adequately capture instructional effectiveness or effort, relying exclusively on SET data is highly problematic. It is possible and advisable to develop a “culture of engagement” that’s as easy as “how’s it going?” check-ins with students. (There is a more formal and evidence-based foundation for this that actually simplifies the process). In short, students can be given multiple low-stakes/high-yield opportunities to provide evidence of teaching and learning effectiveness.

II. Peer-evaluations ought to be implemented.

a. One of the services CETL provides is peer-evaluation of instruction. Working in collaboration with instructors and (often) department heads, CETL staff arrange and conduct class and instructional observations designed to provide informed evidence of instructional effort and effectiveness. Importantly, they include efficient, effective, and customized options and strategies faculty can use to enhance the teaching and learning experience.

b. CETL’s efforts are research-based and consistently result in (1) immediate improvement in the learning experience, (2) student perceptions of it, (3) the integration of efficient (low-risk, high-yield) teaching strategies that make a difference and scale into other classes, and (4) enhanced student and faculty satisfaction. They also provide faculty and department heads with clear evidence of effort to improve student engagement, learning, and teaching.

c. This process supports and unifies the formative development of the instructor/class and the summative/administrative efforts of department heads in a way that is ethical, efficient, informed, and impactful.

d. Some of the documentation, strategies, and forms governing POI (peer observation of instruction) and SGIDs (small group instructional development) are available on the CETL website at https://www.webpages.uidaho.edu/cetl/peer-observation.htm.

III. Self-evaluation is critical.

a. The “informed loop” of faculty development in teaching and learning draws together evidence from student-evaluations, peer-evaluations, and self-evaluations to create a dynamic view of instructional effort and effectiveness.

b. CETL provides the faculty development expertise essential to effective critical self-reflection, the articulation of meaningful teaching
statements/philosophies, and a compelling dossier of evidence in pursuit of excellence in teaching and learning.

The important take-away is that faculty need to be given the opportunity to effectively represent their instructional effort and for it to be complemented, understood, and supported at multiple levels. This requires multiple forms of evidence, and CETL has the expertise and resources to help.

(3) Summary-Review of Research

List of studies referenced, both those referred to TeAC by Terry Grieb and from the faculty member in CLASS, and also additional articles:


Boring, Anne, Kellie Ottoboni, and Philip B. Stark. “Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness.” ScienceOpen Research (January 5, 2016)


Martin, Lisa L. “Gender, Teaching Evaluations, and Professional Success in Political Science.” The Teacher (American Political Science Association, 2016) [relies to a significant degree on citing MacNell and Boring]


Ray, Victor. “Is Gender Bias an Intended Feature of Teaching Evaluations?” Inside Higher Ed, February 9, 2018. [This opinion piece provides no direct citations of sources of claims/research, aside from a mention of Katherine Weisshaar’s study on tenure committees.]

Ryalls, Ken, Steve Benton, Jason Barr, and Dan Li. “IDEA Editorial Note #2: Response to ‘Bias Against Female Instructors” [on Colleen Flaherty’s Inside Higher Education review of Boring et al.]

doi:10.1371/journal.pone.0150194

**Sampling of Review of Research and Responses; see especially Dale Pietzrak’s assessment and comments.**

In a response to Colleen Flaherty’s “Bias Against Female Instructors” *Inside Higher Education* review article (which cites the article by Boring et al.), Ryalls et al. from The IDEA Center (for comparative research data on higher education) state that "Many of the claims made in the IHE article are flagrantly misleading . . . [and that contrary to Flaherty's conclusion that] ‘the association between evaluations and perceived instructor gender in both the U.S. and French data sets is largely statistically significant,’ In point of fact, the correlations between average student evaluations of teaching (SET) scores and gender in the French sample of 1,194 students across six courses ranged from .04 to .11, *none significant at the .05 level*. So, **gender explained no more than 1% of the variance in the ratings.** Putting aside Flaherty's missteps, we have several concerns [list of eight concerns] about the Boring et al. study itself [including this comment “Many of the correlations reported are very weak and non-significant” (Ryalls et al.).

Helen Joyner (TeAC): I went through the actual research articles quickly, plus one that was talked about on *Inside Higher Ed* today and they have some methodology issues. Most concerning is the low n. The IHE article I linked (which is about the Cambridge study you linked) has an n of 2 instructors. The Stanford article has an n of 8. I think these studies are not considering what their actual experimental unit is. They have tons of subreplicates (student responses), but their actual experimental unit is the class, instructor, or name. The comments after the IHE article are a good discussion of some of the issues with these studies. While there may be something worth looking into about bias, we need to be particular about what studies we use as evidence; many of them have significant methodological flaws and should not be used as support for or against a given argument.

From Dale Pietzrak’s “Initial Examination of New Teaching Form Executive Summary 2017”:

“The initial review of the suggested changes to the student feedback on teaching and course was done using results from the Fall of 2016 and Spring of 2017 semesters. This represented 4,718 courses with responses from 33,186 students. Of these, 2,353 courses had 5 or more respondents (28,028 student responses). The numbers for each analysis may vary somewhat based on missing data (i.e., student not responding to items, etc.) . . .
Examination of the relationship of the instructor and course scores (old and new) to class size, number of respondents and response rate was examined.

Examination of the instructor and course scores with instructor gender, citizenship and minority status found no relationship to gender or minority status [Stephan Flores's note: see Slide 31 of Dale Pietrzak’s Powerpoint “Validation Study of New Teaching Questions and Form”, which shows -0.01-0.02 correlations]. A slight decrease in scores was noted for instructors who were not US citizens (about 1.5% shared variance) (pp.1-2).

Follow up from Dale Pietrzak [March 21 to TeAC]: As the gender question(s) seem to remain feeling left unanswered as yet, I went back and did a re-examination to be sure I had not missed something. I re-ran the process using both relationships (point-biserial correlations) and differences (t-tests) and have the results attached. This includes over 3900 courses so we are not lacking the power to detect something in this case. For those in the know, yes, I realize these are essentially mathematically equivalent processes and thus it was somewhat redundant. Sorry. I thought both may be of assistance to illustrate the process. As previously there were no significant differences or relationships when run for all courses, or by any other course level.

Appended/copied below: series of exchanges prior to the February 28th Teaching & Advising Committee meeting, including two different requests to Faculty Senate to explore the issue of gender-inflected (and also other forms of) bias in students’ evaluations of teaching, and Dale Pietrzak’s comments on specific studies as well as summary of our own UI data from recent analyses of transition from older to new Student Evaluation of Teaching (SET) forms.

From: Pietrzak, Dale (dalepietrzak@uidaho.edu)  Sent: Tuesday, February 27, 2018 9:00 AM  
Subject: RE: TeAC: question of Gender Equity in Students' Evaluations of Teaching/mtg. Wed.Feb. 28th
Stephan and all,

While this may seem long, it is really a relatively hasty note. So I apologize for any issues with the text or formatting. I appreciate the discussion and the desire to continue to examine the performance of our SET form. It is a process that should never be taken for granted and needs to continually be reviewed. I also apologize if I was not clear in presenting any of the finding regarding the research and findings with our SET tool, but welcome an opportunity to discuss it further should you desire. As you probably know, measurement and testing were a key research and teaching areas for me as a faculty member, and key in my publications and presentations as I moved from assistant to full professor. For a time this specific area was actually a research focus of mine, both as a faculty member and as I transitioned into my role in university research/assessment. At this time while I keep a somewhat superficial eye on the literature it is certainly no longer a primary research focus area at this time.

Note that one of the key documents/manuscripts presented by Stephan's materials has been reviewed pretty widely, including by one of the most widely used and researched SET tools. Their examination of it is discussed at: [http://www.ideaedu.org/Portals/0/Uploads/Documents/Response_to_Bias_Against_Female_Instructors.pdf](http://www.ideaedu.org/Portals/0/Uploads/Documents/Response_to_Bias_Against_Female_Instructors.pdf). [Flores/Note, citation from the study: “So, gender explained no more than 1% of the variance in ratings.” (Ryalls et al.)] It appears generally consistent with my independent review (below).

However, a few of things need to be noted in a discussion like this as they are very important conceptually.

1.) If you wish to understand the operating characteristic of an instrument you do not give a different instrument to a different population and anticipate that the findings will be similar. If I wanted to understand how the Basic Personality Inventory (BPI) operates with a certain population I would not administer the MMPI-2 to a different population to make statements about the BPI. The results of research, with other tools, can highlight areas where investigations may wish to look, but they do not directly generalize to another tool.

2.) As noted in the PowerPoint and paper resulting addressing the research into our SET, the SET form is not intended to be used in isolation, and should be used as part of a more holistic process to examine both teaching and course structure. There is an expectation that each College has an established teaching and course review process, of which this is to be one part. IEA recommends that this process include other evaluation procedures for teaching and course assessment to supplement and provide context for the results of this tool.

3.) There are few areas in higher education that have the emotional loading that a student evaluation of teaching carries. The only context I have worked where the emotion is as high is when I was testing young children for possible cognitive impairments. Parents are not looking for such information, and as a parent of a delayed child, I understand to some extent why. Similarly, nearly every faculty
member I have met believes himself/herself to be an excellent instructor with well-run and interesting courses. There are few “average” self-ratings here.

4.) Practically every instructor can locate a student in a course to provide support for them. Those wishing to find negative support can generally find a student to support that view as well. Similarly, there has been enough research in this area that one can find some support for nearly any position. Much like student support, one can also find the support of an opposing view in the literature as well. It is key to understanding how our form operates with our population(s).

5.) As was noted in some of the material Stephan sent previously, there are differences in how gender and other characteristics are viewed. It is important to note however, that there are great variations in this by region in the USA and across international cultures. This make it difficult to generalize across cultures in these cases. How my children (minority/disabled), my wife and our family have been treated has varied widely based on where we lived. The locale greatly impacted gender and racial attitudes and treatment when living in urban Ohio (Akron/Cleveland area), northern Kentucky, Miami, Florida, South Dakota and here even here in Idaho. The gender attitudes across these areas is very different in our experience. These regional variations are present in other countries and further complicate the research between countries.

6.) As was noted in the research paper on our SET I provided previously, research never “arrives” at an end with “validity”. Rather it is a journey with the “information to date” being the best we have to work with at any given time.

7.) As expected with any class and instructor there are variations across people and time. So I would expect that what happens generally with a SET form can, and will, see some classes that operate outside of these general operational processes. This is part of why the forms should be used in a context with other information about teaching and classrooms.

8.) There appears to be a great deal of misunderstanding about the word “bias” in the area of assessment. Bias is a form in noise that affects the validity of a tool. Means differences, in and of itself, does not indicate bias. Means differences suggest the need for possible separate scaling of a tool across the groups (gender in this discussion) but it does not mean the tool measures something different between the groups. Bias is essentially the measurement of construct “A” in one group while tapping construct “B” in another group. Scaling differences alone would not typically be an issue of “bias” and is why many tools have gender specific norms.

9.) There are other things that are of importance in this discussion. I am willing to discuss the area as you desire.

Q. What do these tools, and our SET in particular, seem to measure?

A. It appears they measure - The classroom amalgam of student’s perception of how helpful the instructor was in assisting them learn the material. (Educational atmosphere of the classroom).

- The forms do not generally measure grades, effectiveness, or other such aspects (See attached). As such it is one source of information about one aspect of the teaching process.
Again, why the results should be used holistically along with other information.

Now, I have not systematically reviewed all of the papers sent by Stephan at this point. I looked briefly at the first three. If I have time I will look over the others. However, those first three fit within what I have seen previously in the literature and are consistent with what I indicated previously. I have done so in a cursory fashion and have some thoughts about what these papers at this level of review. I certainly could have missed something important.

1.) As I read it, the key finding from the Mengel, Sauermann and Zolitz paper appeared to be, “lower teaching evaluations of female faculty stem mostly from male students, who evaluate their female instructors 21% of a standard deviation worse than their male instructors. While female students were found to rate female instructors about 8% of a standard deviation lower than male instructors.” This means that essentially the effect size was $d=.2$ for gender in this case, this traditionally considered small (http://rpsychologist.com/d3/cohend/). [Flores/Note citation from this study: “Interestingly, we do not observe this gender bias for more senior female instructors or professors” (Mengle et al. 2).] It is of import but does not negates the utility of the tool given the other dimensions of the tool. The tool is very different in construction than what we appear to be using, in addition to the differences in culture by gender that may play a role. As a reviewer for a measurement journal, and associate editor for quantitative methods on another journal I am surprised the value-laden language used in the paper was considered acceptable. As mentioned above, even if these were considered substantive differences in the scores by gender it would suggest the need for separate gender norms or score examination and not that the tool was not something different across gender.

2.) In examining the Boring, Ottoboni and Stark paper, “The numbers show that SET scores do not measure teaching effectiveness well...” These forms are intended to measure student perception of the helpfulness of a teacher relative to their learning the material. There was great variation of the SET to general effects/relationship (table 3) with the overall being 0 (no effect). The “shared variance” of gender and set ranges from 0.0% to about 1.2%. Meaning that 98.8% of the variance was related to other factors. Also, should there have been differences it would not invalidate the tool but rather suggest the need for separate gender examination of results.

3.) MacNell, Driscoll and Hunt had a very small sample. They actually note in their review of the literature, “Studies of gender bias in student ratings of instruction have presented complicated and sometimes contradictory results. Sometimes men received significantly higher ratings (Basow & Silberg, 1987; Sidanius & Crane, 1989), sometimes women (Bachen et al., 1999; Rowden & Carlson, 1996), and sometimes neither (Centra & Gaubatz, 2000; Feldman, 1993). Though they used random assignment it is not clear how they controlled for the impact of gender stereotyping of the instructors. I am sure you remember the movies *Tootsie, Mrs.*
"Doubtfire, Yentl and Madea. Often when asked to “pretend” to be another gender/race/etc. there are unconscious stereotypical behaviors people enact. It was not clear to me from my cursory reading this was addressed. Further, the results reported in the discussion “The difference between the right two columns is significant to the p<=0.10 level.” This is not typically considered statistically significant (p < .05 is the typically accepted lower bound). At best this is a small effect but questionable. Even if this were the case, the outcome would suggest perhaps the desire for separate gender examination of the results not the tool is invalid. The structural equation model (actually redundant and typically not done when PCA is run on the same sample) suggest that the tool is pretty consistent in what the dominant underlying construct is, and it does not appear to be gender.

Note I have no real investment in our SET tool per se. Nor do I really have an investment in any tool. A SET tool not being implemented would free up personnel time and other resources in my office. Additionally I agree, that the questions being raised are important, and need to be examined. The research is never done with any measure and we need to be careful in our use of any tools. As I noted in our findings using our updated tool there is a need for continued research. However, the data we have to date on our SET do not suggest that this is a significant issue with our form/items.

Dale

Dale Pietrzak, Ed.D., LPC-MH(SD), CCMHC
Director Institutional Effectiveness & Accreditation
NWCCU Accreditation Liaison Officer
Administration Building, Room 201
875 Perimeter Drive MS 3163 Moscow, ID 83844-3163
Other: Asso. Editor Quantitative Research JCD

From: "Flores, Stephan (sflores@uidaho.edu)" <sflores@uidaho.edu>
Subject: RE: Validation Study of transition to revised teaching evaluation forms
Date: February 19, 2018 at 5:15:37 PM PST
To: "Flores, Stephan (sflores@uidaho.edu)" <sflores@uidaho.edu>

Stephan,

I had previously investigated the old items (see old attached). I ALSO examined the “new” items separately (see new attached). I summarize the results in the notes of the various slides in the PowerPoint attached for most every examination for reference if you need it. So the relationship of the gender, citizenship and minority status was examined for the new items (along with a number of others things as noted in the report). The notes on slide 31 summarize the findings you are looking for here. I included both the finding from old and new items in the current validation sample as a means to ensure and document consistency of old item examination across samples to show that the results here were not
due to sample specific factors. So there is an examination, it is consistent with that found with the old items and consistent with expectations from previous research in this area.

In sum, neither the new items, nor our old items appeared to show a differential effects/operation between genders. Also, I might suggest, the literature is not really as clear on that issue as your note suggests. Yes, there has been some of the literature suggested the “possibility” existed. No single study stands on its own. In this case the literature has not been a consistent finding, and often the studies that suggest the possibility of this have had issues. Even in the cases where a possible relationship was suggested, this relationship has overwhelmingly been very small and of little to no “practical significance” regarding for the results. See the literature review beginning with slide 3 and running to slide 7 for further information.

If you wish to meet and go over what was done I am willing to do so. Just let me know.

Dale

---------

Copies of separate emails from a professor in CLASS and from Terrance Grieb, sent to Faculty Senate and referred to TeAC:
[prior email of Feb. 19]:
Hello TeAC,
...
This email contains TWO copies of requests for TeAC to look into the issue (question/research) of whether students’ evaluations of instructors have been shown to indicate gender bias.

You will see that the FIRST email copied below is forwarded by Faculty Senate member Jodie Nicotra on behalf of/from a CLASS faculty member.

Below that message is the SECOND email of this message, but a query/request that was forwarded to me by Faculty Senate leadership over the December holiday break, on behalf of Faculty Senate member Terry Grieb (College of Business), including an attached PDF article on gender bias.

I shall send another email to you with some of the brief exchange that I’ve had thus far, with Dale Pietrzak, reminding us of his studies on our teaching evaluation forms, and what they show or don’t show about this question. . . .

Sincerely,

Stephan

From [CLASS faculty member] Date: Tuesday, February 13, 2018 at 8:09 AM To: "Nicotra, Jodie (jnicotra@uidaho.edu)" <jnicotra@uidaho.edu> Subject: Faculty Senate issue/query
Hi Jodie,

My apologies for interrupting your inbox, if this isn’t the normal course. But I don’t really know how the Faculty Senate works. I’m assuming if there is an issue, we raise it with our college senators. You are the one I know best—which is not well. If I’ve misconstrued the process, I am teachable and invite you to let me know how to do this better.

I’ve long been bothered by student evaluations of teaching as an instrument that plays any role in evaluating faculty, either annually or for tenure/promotion. They are deeply problematic, especially for women and for faculty of color (men and women), and I think everyone knows this. This opinion piece (linked below) lays out a lot of my own objections fairly well and includes links to studies. It isn’t hard to find further social science research that demonstrates the ways bias is embedded within these measures. What the opinion piece suggests--perhaps conspiratorially, perhaps not--is that maintaining that bias and the status quo that comes from that very well might be the reason we keep using a measure that all of us know is biased against women and faculty of color while also being a horrible survey instrument from a social science perspective. In my view, there is nothing about them that should be any part of the annual or tenure/promotion review process. Keeping them as part of that system seems irresponsible. Doing so, we knowingly (or with willful ignorance) perpetuate a structure of bias that has the potential to affect people’s livelihoods, not to mention simply their well-being.

Here’s the most recent piece I’ve seen: https://www.insidehighered.com/advice/2018/02/09/teaching-evaluations-are-often-used-confirm-worst-stereotypes-about-women-faculty. It would take two minutes to find peer-reviewed studies that bear out the criticisms this opinion piece charges.

I raised this with my department and got a ho-hum response. In fact, the response was much as the opinion piece would suggest—those protected by the status quo thought the evaluations were a terrific measure and it was pointed out they can be used to dismiss faculty by providing evidence (i.e., justification). I raised it with Andy Kersten, who indicated this is a governance issue that the Senate might take up. So, I’m sharing my concern with you. To be clear, the issue I have isn’t personal. I put no stock in the evaluations and don’t have to, since I’m promoted as far as I can be (and score at or above average). The issue I have is for my colleagues of color and my counterparts who do not possess privileges maleness provides. And, lastly, I’m concerned that an institution based at its core on rational investigation would use something so demonstrably flawed as a tool for evaluation. We might as well use a chicken’s entrails. (Well, perhaps they aren’t that bad.)

Thanks for your time and any attention you might give this and direct the Senate toward. [...]
Miranda (mirandaa@uidaho.edu) <mirandaa@uidaho.edu> Subject: Gender equity in course evaluations

Patrick and Miranda,
...

I am writing to bring up an issue for possible consideration next semester. A faculty member in the CBE raised the issue of gender equality in teaching evaluations. A statement summarizing the issue and supporting research are as follows:

Summary: Several recent empirical studies show that student teaching evaluations are consistently biased against female instructors/faculty across several disciplines and countries. These findings have significant implications on academic hiring, promotion and tenure decisions, as well as the distribution of academic rewards and incentives among faculty. The more universities rely on measures and metrics reflecting gender bias, the further they perpetuate systemic gender bias in academic institutions. In line with its values regarding diversity and inclusion, the University of Idaho needs to take steps to actively reduce or eliminate gender bias from academic evaluation processes, by reducing reliance on measures and metrics that disproportionately and demonstrably disadvantage female faculty.

A large-scale study (N = 19,952) examining this issue was recently published in the Journal of the European Economic Association (please see attached file). The article in The Economist that is discussed in the JEEA study can be located at: https://www.economist.com/news/science-and-technology/21729426-how-long-does-prejudice-last-research-suggests-students-are-biased-against?frsc=dg|e

Additional references on this issue include:


Storage et al. (2016) http://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0150194

MacNell et al. (2014) https://link.springer.com/article/10.1007/s10755-014-9313-4

Here’s an even broader list of references for gender bias in academia: http://blogs.lse.ac.uk/impactofsocialsciences/2016/03/08/gender-bias-in-academe-an-annotated-bibliography/

This might be a topic we could refer to Faculty Affairs. After discussing the issue with my colleague, we also think that this might benefit from some research by Institutional Effectiveness and Accreditation to generate data that might inform a policy decision. It is our position that any policy recommendations on this issue should 1. reflect institution specific data, 2. provide for teaching evaluation adjustments that create a defendable equity policy, and 3. be flexible over time to accurately reflect bias levels within a specific
semester and changing degrees of bias over time. That being said, I believe that there are faculty within the CBE who would be willing to work with Faculty Affairs and/or IEA to research this topic and bring forward results and recommendations for consideration.

I know that nothing will be happening between now and next semester even if it is deemed appropriate to consider this issue, but I wanted to put it out there for your thoughts and comments.

Thank you again and I hope you have a great end of the semester and holiday season.
Best Regards,
Terry

Terrance Grieb, Ph.D.
Professor of Finance
College of Business and Economics
<MengelSauermannZoelitz_GenderBias-Journal of the European Economic Assoc.pdf>