ASSOCIATED COLLEGES OF THE SOUTH

Concerns about Student Evaluations of Teaching (SETs)

The following research was conducted in Spring 2021 to gather information about types of SETs administered at ACS schools, the ways in which SET data are used, and concerns about their current instruments:

- A survey was distributed to each ACS campus to be completed by an administrator (provost, associate dean, or the like) and a faculty member who serves on the review/promotion or tenure/personnel committee, preferably the committee chair.
- Focus groups were also conducted with multiple constituencies, including students, provosts/deans, and members of faculty personnel committees.
- The extensive existing scholarship on SETs was reviewed.

This research revealed six dominant concerns, each of which is described below:

1. **Low response rates:** Given the recent transition at most institutions to online course evaluations that can be completed outside of class, low response rates can skew results. An AAUP survey that had over nine thousand faculty respondents reveals that the student response rate has fallen sharply, "from 80 percent or higher on paper to 20 to 40 percent online" (Vasey and Carroll, 2016). This raises questions about the reliability of the instruments themselves. As Berkeley scholars Stark and Freishtat (2014) point out:

   > The lower the response rate, the less representative the responses might be. There is no justification for assuming that nonresponders are just like responders. Indeed, there is reason to think they are not: They were not present or they chose not to fill out the evaluation. Moreover, people tend to be motivated to act (e.g., fill out an online evaluation) more by anger than by satisfaction.

2. **No faculty input in construction of questions:** Respondents to the AAUP survey also note that they had almost no input in the types of questions asked in student evaluations. Schools that use standardized questions may be "inappropriate [because they] treat all teaching in every field or all students as if they were the same" (Vasey and Carroll, 2016).

3. **Lack of agreement on how to measure "teaching effectiveness":** There is no shared understanding of teaching effectiveness that SETs might help us measure. In the absence of such an understanding, it is unclear what specifically the instruments are helping us measure. Weiman sees this as a shortcoming because "[p]eople are poor at evaluating their own learning, because it is difficult to know what you do not know" (Weiman, 2015). Moreover, students may inaccurately report their learning outcomes, as they may not realize how much they have learned until well after the semester has ended.

   This is one of the ways in which students' implicit bias enters into the evaluation process. Stark and Freishtat argue that, unfortunately, responses to broader questions about teaching effectiveness "are strongly influenced by factors unrelated to learning, such as the gender, ethnicity, and attractiveness of the instructor" (Stark and Freishtat, 2014).

4. **Influence of irrelevant or inappropriate factors:** There have been a large number of studies demonstrating the impact of extraneous factors on student evaluations. In the AAUP survey mentioned earlier, 67 percent of respondents said that "student evaluations create upward pressure on grades." This is particularly challenging for untenured or contingent faculty, for whom evaluations may have direct negative impact on renewal or tenure and promotion decisions (Vasey and Carroll, 2016).

A course's context (discipline, whether it's elective or required, level of difficulty, time of day) may also intensify the effects of bias in student evaluations, especially when these contextual factors intersect with a faculty member's identity. For example, a faculty member who teaches a course in which standard pedagogical practice is to "nurture" a student's expressive ability (such as a composition course) may then by confronted with an angry student who *feels* as though the nurturing (read: maternal) faculty member has no authority to judge harshly (i.e., assign a low grade) (Lauer 276).

Focus groups conducted with ACS faculty and students revealed that student evaluations tend toward emotional extremes. Students are motivated to respond affectively either because they really enjoyed a class or instructor or because they had a negative experience. Students reporting on their (emotional) experience in a course, however, may not be useful in determining an instructor's effectiveness as a teacher. In other words, student evaluations are not always constructed and administered in ways that measure teaching effectiveness.

Such emotional responses factor into quantitative data as well as qualitative comments. Narrative comments from students tend to resemble Yelp reviews at best and are often "abusive and bullying" in their tone. Some faculty members shared that "comments from students are on the extremes: those who are very happy with their experience or grade, and those who are very unhappy." More significantly, "Some women faculty members and faculty members of color report receiving negative comments on appearance and qualifications; it seems that anonymity may encourage such inappropriate and sometimes overtly discriminatory comments" (Vasey and Carroll, 2016).

Student evaluations influenced by these factors are unhelpful in terms of offering guidance for improvement. Weiman reports that students do not (or cannot) offer clear direction on ways to improve; "there is typically a distribution of strongly felt positive and negative student opinions on nearly every aspect of the teaching," making it difficult for faculty to figure out which advice to follow (Weiman, 2015).

5. **Misleading comparisons among instructors:** SETs inadvertently ask students to evaluate instructors based on previous experience. Weiman argues that students "judge the effectiveness of an instructional practice . . . by comparing it with others that they have already experienced" (Weiman, 2015). That is, students who have mostly taken large lectures may not be in a position to appropriately evaluate a discussion-based seminar. This may also discourage faculty from using innovative teaching methods.

   Students are not the only ones who rely on such comparisons; SETs that offer mean or average data for faculty in the department, discipline, or institution also encourage such comparisons. Stark and Freishtat argue against using departmental averages, as it is impossible to tell whether someone's lower-than-average score is actually concerning, "because of instructor-to-instructor and semester-to-semester variability."

6. **Bias:** This research revealed that a common thread running through all of these concerns is the issue of implicit bias in SETs. High-stakes decision-making about a faculty member's tenure, promotion, merit, and awards is often made by consulting evaluations that, by design, perpetuate biased responses related to the race, gender, age, nationality, and/or sexual orientation of the faculty member being evaluated.

   Studies demonstrating the existence of implicit bias in student evaluations goes back to at least the 1970s. Again and again, scholarship shows evidence of how students' perceptions and stereotypes regarding race, gender, ethnicity, sexual orientation, and national identity significantly impact how they rate professors' teaching effectiveness. Here are but a few examples:
   - In a study with 87 students participants (41 male and 46 female), students rated a lecture they perceived to have been authored by a male professor as "significantly higher" than the same lecture perceived to be authored by a female professor (Abel and Meltzer).
   - Miller and Chamberlain (2000) show that students call women faculty "teachers" and male faculty "professors," regardless of their actual rank.
   - MacNeill, Driscoll, and Hunt (2015) report that, in an online experiment, students rated instructors perceived as female lower than those perceived as male.

- Mengel, Sauermann, and Zölitz (2017) analyzed 20,000 student evaluations at a university in the Netherlands and found that female instructors scored 37 percentage points lower than male instructors. The bias is most evident in male students and affects junior female faculty the most.
- Chávez and Mitchell conducted a study of 14 online sections of two Political Science courses at a large public university. They found more negative comments for female instructors, implying that "there is a noticeable difference between the types of comments that women and faculty of color received" (2019). For example, here is a comment they cite for a white female professor: "She got super annoyed when people would email her and did not come off as very approachable or helpful."
- Sprague and Massoni (2005) offer a comprehensive analysis of how gendered expectations influence students' qualitative feedback on SETs.