> **CALIFORNIA STATE UNIVERSITY FULLERTON**
> **ACADEMIC SENATE**
> **SOQ COMMITTEE FINAL REPORT**
> **SPRING 2019**

INTRODUCTION

The SOQ Committee was appointed by the CSUF Academic Senate as a result of a Response to a Statement of Opinion election held in Spring 2017, in which the campus electorate voted in support of the following proposal:

> An ad hoc committee should be formed to review the research literature on student opinion questionnaires, compare their findings to current campus policy, and make evidence-based recommendations for changes to our SOQ policies.

Therefore, the SOQ Committee's specific charge was to "Review the research literature on student opinion questionnaires, compare the findings to current campus policy, and make evidence-based recommendations for changes to CSUF's SOQ [1]policies." The committee members were faculty drawn from each college: Eliza Noh (HUM), Cynthia King (COMM), Lidia Nuño (SOC SCI), Catherine Brennan (NSM), Patrice Waller (EDU), Hope Weiss (ECS), Peggy Shoar (HHD), Marc Dickey (ARTS), and Mira Farka (MCBE). Ed Collom (FAR) served as an ex officio member of the committee. The committee divided its work among separate subcommittees in order to examine two main areas: 1) bias and validity among SOQs and 2) SOQs implemented online and SOQs among online v. face-to-face instruction.

This report represents the culmination of one semester of research and discussion among the committee members. It is intended as a jumping off point for what must involve a campus-wide discussion among CSUF stakeholders. The SOQ committee itself did not share a consensus on some of its recommendations, which indicates the need for further examination and discussion. The report is structured according to the committee charge: 1) literature review, 2) campus policy review, and 3) recommendations.

LITERATURE REVIEW

Overall Validity & Reliability of SETs: What Do or Don't They Measure?

---

[1] Within the research literature, evaluations of teaching administered to students are usually referred to as "student evaluations of teaching" (SETs). While CSUF previously called student evaluations of teaching performance "student ratings of instruction" (SRIs), they are currently referred to as "student opinion questionnaires" (SOQs). Since campus policies (e.g., CBA and UPS) use SOQs as student evaluations of teaching performance, SOQs and SETs will be referred to interchangeably for the purposes of this report.

- The structure of teaching quality is multidimensional. Research tested the reliability of a "students' opinion questionnaire" (SOQ) and "student's evaluation of education quality questionnaire" (SEEQ). The results showed that although both questionnaires had acceptable reliability, the SEEQ better revealed teaching's multiple dimensions and also confirmed the Marsh opinion that believed that the structure of teaching quality is multi-dimensional (Marsh & Dunkin, 1992; Rezaei, Ghartappeh, Bagher Kajbaf, Safari, Mohammadi, & Sharafi, 2018).

- SETs, in particular, omnibus "teaching effectiveness" measures, are poor measures of teacher effectiveness, operationalized as student performance in current or future courses (Boring, Ottoboni & Stark, 2016; Stark & Freishtat, 2012; Wigington, Tollefson & Rodriguez, 1989; Uttl, White, & Gonzalez, 2017; Williams, 2007).

- The interactions between instructor and class variables need to be considered when universities use student ratings of instruction to evaluate faculty. It would be unwise to assume that the same numerical rating means the same level of effectiveness when it is earned by instructors of different sexes at different professional ranks, teaching classes at different educational levels, and using different instructional formats. Class level, class size, and type of instruction, along with instructor reputation, instructor rank, and instructor sex need to be included in this consideration (Wigington, Tollefson & Rodriguez, 1989).

- SETs are more sensitive to students' gender bias and grade expectations than they are to teaching effectiveness (Boring, Ottoboni & Stark, 2016).

- Teacher effectiveness as measured by students' performance on end-of-semester exams was negatively correlated with teacher effectiveness as measured by performance in subsequent courses: "Just as it is misguided to assume that ratings have any obvious relationship with student learning, it is also misguided to assume that end-of semester test performance is a valid measure of deep learning" (Braga, Paccagnella & Pellizzari, 2014, p. 85).

- There is some debate on the relationship between SET ratings and student outcomes based on meta-analyses. One study found that using large sample sizes showed no or minimal correlation between SET ratings and learning outcomes (Uttl, White & Gonzalez, 2017, p. 22—available online 2016). But, this interpretation has been challenged by Ryalls, Benton & Li (2016), who conclude that SET validity is better supported than any other teaching evaluation strategy.

- An evaluation often tells more about a student's opinion of a professor than about the professor's teaching effectiveness (Williams, 2007).

- SETs fail to measure important dimensions of teaching quality (Fox & Keeter, 1996).

- The use of student ratings in academic trajectories should be used with caution considering the biases they account for (MacNell, Driscoll & Hunt, 2014).

- SETs can provide useful insights for assessing teachers. In their review of the literature, Kornell and Hausman (2016) conclude that, although students may not have the expertise to recognize good teaching, "their reports reflect their experiences, including whether they enjoyed the class, whether the instructor helped them appreciate the material, and whether the instructor made them more likely to take a related follow-up course. We think that these factors should be taken into account when assessing how good a teacher is" (p. 7).

- Stark and Freishtat (2014) found that "Student ratings of teaching are valuable when they ask the right questions, report response rates and score distributions, and are balanced by a variety of other sources and methods to evaluate teaching (p.1). They also reference Laurer (2012) in stating that "Students are in a good position to observe some aspects of teaching, such as clarity, pace, legibility, audibility, and their own excitement (or boredom). SETs can measure these things" (p. 4).

- As cited in Linse (2017), SET validity has been tested more than any other method for evaluating faculty teaching (Abrami, 2001; Abrami, d'Apollonia & Cohen, 1990; Aleamoni, 1999; d'Appolonia & Abrami, 1997; Feldman, 1989; Marsh, 1982b, 1984; Marsh & Roche, 1997). Citing Berk (2005, 2013a) and McKeachie (1997), Linse states that "the majority of the legitimate research on student ratings indicates that they are a more reliable and valid representation of teaching quality than any other method of evaluating teaching, including peer observation, focus groups, and external review of materials" and that "they are highly correlated with other measures of teaching effectiveness" (Abrami, d'Apollonia & Cohen, 1990; Berk, 2013a; qtd. in Linse, 2017, p. 97). However, Linse may be overstating Berk and McKeachie. Berk (2005, 2013a) proposes a unified conceptualization of teaching using multiple sources of evidence, such as student ratings, peer ratings, and self-evaluation, to provide an accurate and reliable base for formative and summative decisions. Further, Berk (2013a) concludes that higher education in general falls short of the available technology to comprehensively assess teaching effectiveness. McKeachie (1997) believes that student ratings are the single most valid source of data on teaching effectiveness but adds, "However, student ratings are not perfectly correlated with student learning, even in the validity studies carried out in large courses with multiple sections" (p. 1219), and argues that the problem lies neither in the ratings nor in the correction but rather in the lack of sophistication of personnel committees who use the ratings.

- Global items in student ratings of instruction, or a general broad-stroke, summary index of teaching performance or course quality, can be unreliable, unrepresentative of the domain of teaching behaviors it was intended to measure, and inappropriate for personnel decisions according to US professional and legal standards (Berk, 2013b).

Gender Bias

- SETs are biased against women (Basow & Silberg, 1987; Boring, Ottoboni & Stark, 2016; MacNell, Driscoll & Hunt, 2014; Mitchell & Martin, 2018).

- The pattern of gender bias depends on multiple factors. One study shows that the effect is most strongly negative when the instructor is a female graduate student instructor, and disappears when the instructor is a female senior professor. The negative effect against female instructors is also stronger when the content is math-heavy, the class is predominantly male, and the instructor is young. On the other hand, female students give higher ratings to female instructors when the instructor is senior, and such bias can favor female instructors over male instructors when a class is majority female (Mengel, Sauermann & Zolitz, 2018).

- Male professor ratings are unaffected by student gender. However, male students gave female professors lower ratings (Basow, 1995).

- Because gender bias depends on so many factors, it is not possible to adjust for the bias. Gender biases can be large enough to cause more effective instructors to get lower SETs than less effective instructors (Boring, Ottoboni & Stark, 2016).

- 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 (MacNell, Driscoll & Hunt, 2014).

- The language students use in evaluations regarding male professors is significantly different than language used in evaluating female professors. They also show that a male instructor administering an identical online course as a female instructor receives higher ordinal scores in teaching evaluations, even when questions are not instructor-specific. Findings suggest that the relationship between gender and teaching evaluations may indicate that the use of evaluations in employment decisions is discriminatory against women (Mitchell & Martin, 2018).

- Female professors are rated higher in measures of instruction quality—students felt female professors are better teachers and that their class environments promoted significantly more learning (Whitworth, Price & Randall, 2002).

- Use of evaluations in employment and promotion decisions can be discriminatory against women (Mitchell & Martin, 2018). But, they conclude that "Research on this issue is far from complete. Our findings are a critical contribution, but more research is needed before the long-standing tradition of using SETs in employment decisions can be eliminated (p. 652).

Racial Bias

- Racial minority instructors tend to receive significantly lower SET scores compared to white male instructors (Merritt, 2007).

- One study results showed that racial minority faculty, particularly African Americans and Asians, were evaluated more negatively than White faculty in terms of overall quality, helpfulness, and clarity. Black male faculty were rated more negatively than other faculty (Reid, 2010).

- African American faculty scored lower mean scores on SETs than other measured racial/ethnic groups, including White, Asians, Latinos, and Native Americans (Smith & Hawkins, 2011).

- Students perceive faculty of color discussing "controversial issues" pertaining to racial inequity as being biased (Littleford & Jones, 2017).

- Students expect African American professors to be more biased when teaching race/diversity focused topics (Littleford, Ong, Tseng, Milliken & Humy, 2010).

- Ratings of professor warmth and availability for Latino professors appear to be contingent on their teaching style, whereas the rating of Anglo professors' warmth is less contingent upon teaching style. Among women professors with strict teaching styles, Anglo women were rated as more capable than Latinas with the same teaching style. Lenient Latinos were viewed as more capable than strict Latinos, whereas Anglo professors' capability was not contingent on teaching style. Anglo male students, more so than any other group of students, judge women professors and Latino professors as having a political agenda (Anderson & Smith, 2005).

Other Biases: Language, Age, Physical Appearance, Rank, Discipline, Course Level, Student Educational Level

- Professors perceived as attractive received student evaluations about 0.8 of a point higher on a 5-point scale (Riniolo, Johnson, Sherman & Misso, 2006).

- Instructors who are viewed as better looking receive higher instructional ratings, with the impact of a move from the 10th to the 90th percentile of beauty being substantial. This impact is larger for male than for female instructors (Hamermesh & Parker, 2005).

- Students rated the "young" male professor higher than they did the "young" female, "old" male, and "old" female professors on speaking enthusiastically and using a meaningful voice tone during the class lecture regardless of the identical manner in which the material was presented (Arbuckle & Williams, 2003).

- Tenured professors tend to have higher ratings than untenured (Marsh & Dunkin, 1992; Feldman, 1983).

- There can be an intersection of bias with implications for tenure and promotion. In her sample, Basow (1995) found that females are less likely to be tenured or promoted to professor (see also Basow & Silberg, 1997).

- Women with junior faculty status tend to score lower—male student respondents typically drive these lower scores. This sometimes results in female faculty spending even more time on "improving" their teaching and keeps them away from conducting research (Mengel, Sauermann & Zolitz, 2018).

- Humanities have higher ratings than sciences (Marsh & Dunkin, 1992; Feldman, 1983).

- Upper division courses are rated higher than lower division (Marsh & Dunkin, 1992; Feldman, 1983).

- Graduate students give higher scores than undergraduate students (Whitworth, Price & Randall, 2002).

- Non-native English speakers are more likely to receive lower SET scores. When examining the interaction between gender and non-native English, women that were non- native English speakers had the lowest scores (Fan et al., 2019).

- Chang, Zhang and Chen (2012) offer evidence of cultural bias in common SET questions. They advocate for changes that include the use of more interculturally and globally sensitive instruments.

In-Person vs. Online Evaluations/Evaluations of In-Person vs. Online Instruction

- In terms of administration, electronic evaluations are a more cost effective, efficient and reliable data collection strategy for student evaluations than paper evaluations (Collom & Calucag, 2019; Standish, Joines & Gallagher, 2018).

- Study results find that response rates are lower for online evaluations (Mau & Opengart, 2012; Standish, Joines, Young & Gallagher, 2018) but that they can be improved:
    - By dedicating course time to complete them (Nevo, McClean & Nevo, 2010; Standish, Joines, Young & Gallagher, 2018).
    - When the instructor assures students that their evaluations are valued (Chapman & Joines, 2017).
    - By insuring a wide enough time window for asynchronous responses. Online, higher performing students tend to submit earlier and more positive evaluations, while lower performing students tend to submit their evaluations later (Estelami, 2015).

- Recent studies find negligible differences in aggregate scores of evaluation ratings administered online versus in person when other biases and conditions are accounted for (Kelly, Ponton & Rovai, 2007; Risquez, Vaughan & Murphy, 2015; Standish, Joines, Young & Gallagher, 2018; Stanny, Gurung & Landrum, 2017).

- Results of one study found lower racial bias for online asynchronous evaluations (Carle, 2009).

- There are notable qualitative differences in the evaluations of online and face-to-face courses, which were considered to reflect differences in instruction rather than bias in evaluations (Kelly, Ponton & Rovai, 2007).

REVIEW OF CAMPUS POLICIES AND PRACTICES

A critical review of student evaluations is not a new endeavor in the CSU. In 2008, a joint committee of the California State University (CSU), the California Faculty Association (CFA), and the CSU Academic Senate (CSUAS) produced a report on "the best and most effective practices for the student evaluation of faculty teaching effectiveness" (p. 2). The report was based on a survey of twenty-two participating CSU campuses regarding their current student evaluation practices. Campuses reported on how forms were developed and if they included the opportunity for comments and qualitative feedback.

The committee considered the following questions: 1) what do student evaluations measure; 2) what factors influence the results of student evaluations; 3) what are the characteristics of well- designed teaching evaluations; and 4) how can student evaluations be used most effectively? The committee found that student evaluations are not a simple measure of teaching effectiveness and that they present a challenge for evaluating faculty: "student evaluations are 'ratings derived from students' overall *feelings* that arise from an *inseparable* mix of learning, pedagogical approaches, communication skills, and affective factors that may or may not be important to student learning'" (Nuhfer, 1996 qtd. in CSU, CFA & CSUAS, 2008, p. 4, original emphases).
The committee noted student variables that can influence the outcome of evaluations, including student motivation, anticipated grades, and the perceived difficulty of the course (p. 4), as well as online evaluation formats, which resulted in lower response rates and higher likelihood for defamatory or offensive speech that attacks the instructor (p. 6).

Recommendations were made in the following areas: administering evaluations, reporting results, determining which courses to evaluate, determining content and design of evaluation, and making recommendations most effective. Considerable attention was paid to online evaluations, particularly in the areas of security, confidentiality and anonymity, response rates,non-direct comparability of online v. in-person evaluations, and processes for challenging evaluations in personnel files. The committee concluded that student evaluations primarily measure student satisfaction and, as such, should never be the sole basis for evaluation of teaching effectiveness. Further, student evaluations must be recognized as only

one component of an evaluation of teaching effectiveness (p. 9). Recommendations were also made to the following groups: Chancellor's Office, Academic Senate CSU, CFA, Provosts, and campus Academic Senates and faculty. Notably, the committee recommended that the Chancellor's Office examine the influence of racial, ethnic, and gender bias on evaluations (p. 10).

As for current SOQ policy and practices at CSUF, the CFA/CSU Collective Bargaining Agreement (2015) states that periodic evaluation procedures "shall, for tenure-track faculty unit employees who teach, include, but not be limited to, student evaluations of teaching performance." Student evaluations of teaching are also required for the periodic evaluation of full-time temporary faculty unit employees (CFA, p. 52). The CSUF University Policy Statements 210.002, 210.020, and 220.000 pertain to SOQs. In particular, UPS 210.002 states, "Student Opinion Questionnaires contribute significantly to the evaluation of a faculty member's teaching effectiveness" (p. 6), suggesting that SOQs correlate with teaching effectiveness. Per UPS 210.020, consideration of SOQs is required in the periodic evaluation of tenured faculty.

Notably, the 2008 joint committee report concluded that there was no single definition of teaching effectiveness, and, in spite of the report being produced in 2008, UPS 220.00 on Policies, Procedures, and Guidelines for the Administration of SOQ Forms has not been revised since 2007.

In a non-systematic comparison across colleges and departments, the SOQ committee found inconsistent interpretations of UPS pertaining to SOQs in terms of varying departmental personnel standards, SOQ questionnaire forms (with 106 different forms campus-wide), and practices (e.g., how SOQs are administered and analyzed, with different weight placed on numerical scores vs. written comments). Student response rates on SOQs are also widely variable (Collom & Calucag, 2019).

OTHER OBSERVATIONS AND CONSIDERATIONS

- An analysis of CSUF SOQ data to test for bias (instructor age, gender, race, etc.) has not been done. Because it would require coding faculty characteristics (age, gender, race), drawing from a variety of sources, such an analysis was beyond the scope of the committee's work and would potentially compromise policies on confidentiality and privacy.

- Several members of this committee with significant DPC experience noted that they frequently found the written comments of students to be more valuable than the raw scores when evaluating departmental faculty teaching performance. We did not find literature assessing the value of student comments, but we were not able to conduct an exhaustive search.

- SOQs have the potential to expose inappropriate faculty conduct that would not be

apparent via announced peer observations or evaluations of teaching materials. Such inappropriate conduct could include frequent cancellation of class and office hours, ineffective use of class time, frequent change of assignment due dates, failure to return graded material in a timely manner, or favoritism.

- SOQs have the potential to offer valuable feedback to faculty, allowing them to see the student perspective and improve teaching practices.

- In a non-systematic comparison of SOQ forms among CSUF departments, significant variability in the nature, and arguably the quality, of the questions posed to students was noted.

- Similarly, it was discovered that different CSUF departments use a variety of inappropriate ways to incorporate SOQ scores into faculty evaluations: (i) some departments specify SOQ score ranges that qualify teaching as "outstanding," "very good," etc., precise to two decimal places. These are applied irrespective of course format (lecture vs. discussion), level (100-level vs. 400-level), or enrollment (many or few students); (ii) some departments specify that all junior faculty must achieve "above average" SOQ scores in order to earn tenure; and (iii) some departments merely specify that SOQ scores must be above a certain minimum, without categorizing score ranges as "outstanding," "very good," etc.

- Several members of the committee perceived that SOQ scores are often the only quantitative element in an instructor portfolio, and thus requiring the least effort to evaluate, and also giving a false sense of precision. They are often the first element of a teaching portfolio that departmental personnel committees look for. This seems to contribute to a well-studied cognitive bias known as "anchoring bias," in which early information biases decision-making, with later information discounted if it is not in accord with early information (Tversky & Kahnemann, 1974).

- Complete elimination of SOQs would potentially place all evaluation decisions in the hands of departmental faculty on personnel committees, based on peer observation and/or evaluation of teaching materials. Faculty are specialists in their discipline, but perhaps not pedagogy, and may not be better qualified than students to evaluate teaching effectiveness. Moreover, lack of fairness in evaluation of faculty by faculty (whether of teaching or research) stemming from departmental politics, personal grudges, and bias is well-documented. Therefore, while SOQs may be problematic, their elimination has the potential to introduce *more* bias to evaluation of teaching (Ryalls, Benton & Li, 2016).

- Braga, Paccagnella & Pellizzari (2014) found bias but suggest ways SETs might be used differently. For example, they suggest that "since the evaluations of the best students are more aligned with actual teachers' effectiveness, the opinions of the very

good students could be given more weight in the measurement of professors' performance" (p. 85).

- Stark and Freishtat (2014) suggest that reliably and routinely measuring teaching effectiveness will never happen because it does not seem possible to effectively define it. In light of the controversy surrounding the utility of SETs, they have made the following recommendations:
  - Drop omnibus items about "overall teaching effectiveness" and "value of the course" from teaching evaluations.
  - Do not average or compare averages of student rating scores. Instead, report the distribution of scores, along with the number of responders and the response rate.
  - Pay careful attention to student comments—but understand their scope and limitations.
  - Students are the authorities on their experiences in class, but typically are not well situated to evaluate pedagogy generally.
  - When response rates are low, extrapolation is unreliable.
  - Avoid comparing teaching in courses of different types, levels, sizes, functions, or disciplines.
  - Use teaching portfolios as part of the review process.
  - Use classroom observation as part of milestone reviews.
  - To improve teaching and evaluate teaching fairly and honestly, spend more time observing the teaching and looking at teaching materials (p. 6).

- McDaniel (2018) suggests including language in the course syllabus that addresses student evaluations. Such language would increase student awareness of their learning experience throughout the semester and promotes the need for meaningful responses on the evaluation. Consider examples provided on https://s3.amazonaws.com/vu- cft/eval/examples.html

- Some committee members noted that the high number and frequency of portal notifications that students receive reminding them to complete the SOQs could negatively influence students' responses.

- In order to increase response rates, appropriate incentives that do not contribute to response bias could be provided for students to encourage them to complete their SOQs.

RECOMMENDATIONS

- Reconstitute the Academic Senate SOQ committee next year to examine current campus SOQ practices for bias, variability of SOQ practices across campus, and effective alternatives to SOQs.

- Extend committee work to expand dialogue and inquiry on SOQ practices beyond

the CSUF campus.

- FAR should continue to monitor student response rates for online vs. paper SOQs.

- Potential recommendations for future deliberation:
  - ❖ Relevant University Policy Statements should be revised to convey that SOQs are valued indicators of student *opinions* but are not reliable measures of student learning.
  - ❖ Work with CFA to modify CBA articles on student evaluations of teaching performance and align UPS with any CBA changes.
  - ❖ Departments should establish policies for SOQ interpretations that acknowledge their limitations and potential biases.
  - ❖ Departments should establish policies requiring additional, appropriate indicators and criteria for teaching performance that are independent from SOQs.
  - ❖ All SOQs should go online to the extent possible for more reliable, cost and time efficient processing, and to avoid biases created by inconsistencies in administration.
  - ❖ The literature mostly demonstrates that SETs are not a valid measure of teaching effectiveness if conceptualized as student performance or learning. Based on this, either: 1) stop using SOQs as an objective measure of teaching effectiveness and rather as measures of student opinion or perception or 2) deemphasize SOQs as indicators of teaching effectiveness, and rather consider them as one type of data to be used in the overall evaluation of teaching effectiveness (Rezaei, Ghartappeh, Bagher Kajbaf, Safari, Mohammadi, & Sharafi, 2018; Oermann, Conklin, Rushton & Bush, 2018).
  - ❖ The onus should be on universities that rely on SOQs for employment decisions to provide convincing affirmative evidence that such reliance does not have disparate impact on women, underrepresented minorities, or other protected groups. Because the bias varies by course and institution, affirmative evidence needs to be specific to a given course in a given department in a given university. Absent such specific evidence, SOQs should not be used for personnel decisions (Boring, Ottoboni & Stark, 2016, p. 11).
  - ❖ Any decisions stemming from SOQs "should be made very tentatively and alongside other indices of instructional effectiveness such as statements of the instructor's teaching philosophies, duties, and short- and long-term goals and objectives; self- evaluations undertaken by the instructor; evaluations by peers and administrators; unsolicited written comments made by students; samples of students' work; and records of student achievement after leaving the course and/or institution" (Seldin qtd. in Onwuegbuzie, Daniel & Collins, 2009, p. 207).
  - ❖ Analyse existing CSUF SOQ data for the existence of bias according to faculty age, race, gender, etc. If data are not available, collect data to assess this question. Include student study sample.
  - ❖ Use SOQs as a means of obtaining student feedback and opinions that may be considered by the instructor for instructional or course improvement.
  - ❖ Consider the utility of SOQ scores as a measure of improvement of ability to engage students over consecutive semesters in the same course by a single instructor. Keep

in mind that, as measures of student opinion, SOQs are not a measure of student learning outcomes or teaching effectiveness, but may nevertheless provide some useful feedback about the student experience.

❖ When using SOQs as part of the evaluation of student experience of faculty teaching performance, departmental personnel committees should follow evidence-based guidelines and best practices. These include: (i) not relying on fixed SOQ ranges that use the same scale across courses of different levels and formats (e.g,. 200-level vs. 400-level courses, lecture vs. seminar/discussion courses, large vs. small classes); (ii) not averaging into an aggregate score a faculty member's SOQ scores from courses of different formats; and (iii) not requiring that all junior faculty achieve "above average" SOQ scores.

❖ If departmental personnel documents are permitted to specify SOQ score ranges that characterize instruction as "outstanding," "very good," etc., then they should be required to specify also how other measures of teaching effectiveness shall be quantified, including peer evaluations, quality of teaching materials and assessments, self-reflections, etc. This is to avoid the cognitive bias that over-weighs quantitative measures relative to qualitative measures.

❖ Data analysis should not be limited to averages but also consider data distribution and dispersion (Stark & Freishtat, 2014).

❖ Develop guidelines and recommendations for improving the quality of questions posed to students on SOQ forms. This should include examples of useful, validated questions to ask students.

❖ Develop ways to standardize the use of student comments in the evaluation of teaching.

❖ Explore validated alternatives to SOQs, such as peer evaluations of teaching, as measures of teaching effectiveness.

❖ Develop guidelines and training to assist departmental personnel committees in the fair, unbiased, and rigorous assessment of faculty teaching effectiveness during the RTP process.

❖ Consider hiring a permanent educational specialist in the Faculty Development Center to assist departments and faculty in the assessment and improvement of faculty teaching effectiveness.

❖ Consider adopting practices for evaluating and improving faculty teaching used by other institutions, such as offering optional confidential classroom observations and feedback by professional educators (Rice University, https://cte.rice.edu/observations) or developing trainings on pedagogy for new faculty, to be offered along with a 3-unit assigned time in the first semesters.

❖ Set a minimum number of students required for SOQ administration in order to ensure anonymity.

❖ Establish a notification policy that is more streamlined and less burdensome for students in order to remind them to complete the end-of-term SOQs.

❖ Establish appropriate incentives for students to complete the end-of-term SOQs.

REFERENCES

Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. New Directions for Institutional Research, 109, 59–87.

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). The validity of student ratings of instruction: What we know and what we don't. Journal of Educational Psychology, 82(2), 219– 231.

Aleamoni, L. M. (1999). Student rating myths versus research facts: An update. Journal of Personnel Evaluation in Education, 13(2), 153–166.

Anderson, K. J., & Smith, G. (2005). Students' Preconceptions of Professors: Benefits and Barriers According to Ethnicity and Gender. Hispanic Journal of Behavioral Sciences, 27(2), 184–201. doi:10.1177/0739986304273707

Arbuckle J., & Williams B. D. (2003). Students' perceptions of expressiveness: age and gender effects on teacher evaluations. Sex Roles, 49: 507–516. doi.org/10.1023/A:1025832707002

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. Journal of Educational Psychology, 87(4), 656.

Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? Journal of Educational Psychology, 79(3), 308–314. doi:10.1037/0022-0663.79.3.308

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. International Journal of Teaching and Learning in Higher Education, 17(1), 48–62.

Berk, R. A. (2013a). Top 5 flashpoints in the assessment of teaching effectiveness. Medical Teacher, 35(1), 15–26. doi: 10.3109/0142159X. 2012.732247

Berk, R. A. (2013b). Should global items on student rating scales be used for summative decisions? Journal of Faculty Development, 27(1), 57–61.

Boring, A., Ottoboni, K., & Stark, P. (2016). Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness. ScienceOpen Research. doi:10.14293/s2199-1006.1.sor- edu.aetbzc.v1

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. Economics of Education Review, 41, 71–88. doi:10.1016/j.econedurev.2014.04.002

California Faculty Association. (2015). Collective Bargaining Agreement Between the Board of Trustees of the California State University and the California Faculty Association, Unit 3-Faculty, November 12, 2014-June 30, 2017 (Extended to June 30, 2020).

California State University, California Faculty Association, & Academic Senate, CSU. (2008). Report on Student Evaluations of Teaching.

Carle, Adam C. (2009). Evaluating College Students' Evaluations of a Professor's Teaching Effectiveness across Time and Instruction Mode (Online vs. Face-to-Face) Using a Multilevel Growth Modeling Approach. Computers & Education, 53(2), 429-435.

Chang, C., Zhang, M., & Chen, Z. J. (2012). Cultures matter: Alternative model of teaching evaluations. China Media Research, 8(2), 86-93.

Chapman, Diane D., & Joines, Jeffrey A. (2017). Strategies for Increasing Response Rates for Online End-of-Course Evaluations. International Journal of Teaching and Learning in Higher Education, 29(1), 47-60.

Collective Bargaining Agreement between the Board of Trustees of the California State University and the California Faculty Association, Unit 3 Faculty, November 12, 2014 – June 30, 2017 (Extended to June 30, 2020).

Collom, E., & Calucag, N. (2019). Student Opinion Questionnaires: Overview, Processes, and Response Rates [Presentation]. Retrieved from http://www.fullerton.edu/far/soq/.

d'Apollonia, S., & Abrami, P. (1997). Navigating student ratings of instruction. American Psychologist, 52, 1198-1208.Available at ISU Parks Library, General Collection: BF1 Am38

Estelami, H. (2015). The Effects of Survey Timing on Student Evaluation of Teaching Measures Obtained Using Online Surveys. Journal of Marketing Education, 37(1), 54-64.

Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. PloS one, 14(2), e0209749.

Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education*, *18*(1), 3-124.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. Research in Higher Education, 30(2), 137–189.

Fox, J. C., & Keeter, S. (1996). Improving Teaching and Its Evaluation: A Survey of Political Science Departments 1. PS: Political Science & Politics, 29(2), 174-180.

Hamermesh, D.S., & Parker, A. (2005). Beauty in the classroom: instructors pulchritude and putative pedagogical productivity. Econ Educ Rev, 24(4): 369–376.

Kelly, Ponton, & Rovai. (2007). A comparison of student evaluations of teaching between online and face-to-face courses. The Internet and Higher Education, 10(2), 89-101.

Kornell N and Hausman H (2016) Do the Best Teachers Get the Best Ratings? Frontiers in Psychology 7, 570, 1-8.

Linse, A. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. Studies in Educational Evaluation, 54, 94-106.

Littleford, L. N., & Jones, J. A. (2017). Framing and source effects on White college students' reactions to racial inequity information. Cultural Diversity and Ethnic Minority Psychology, 23(1), 143–153. doi:10.1037/cdp0000102

Littleford, L. N., Ong, K. S., Tseng, A., Milliken, J. C., & Humy, S. L. (2010). Perceptions of European American and African American instructors teaching race-focused courses. Journal of Diversity in Higher Education, 3(4), 230.

MacNell, L., Driscoll, A., & Hunt, A. N. (2014). What's in a Name: Exposing Gender Bias in Student Ratings of Teaching. Innovative Higher Education, 40(4), 291–303. doi:10.1007/s10755- 014-9313-4

Marsh, H. W. (1982b). Validity of students' evaluations of college teaching: A multitrait- multimethod analysis. Journal of Educational Psychology, 74, 264–279.

Marsh, H. W. (1984). Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76(5), 707–754.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), Higher education: Handbook of theory and research (vol. 8, pp. 143-233). New York: Agathon Press. Available at ISU Parks Library: LB2324 H54x

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. American Psychologist, 52, 1187–1197.

Mau, Ronald R., & Opengart, Rose A. (2012). Comparing Ratings: In-Class (Paper) vs. out of Class (Online) Student Evaluations. Higher Education Studies, 2(3), 55-68.

McDaniel, R. (2018, August 13). Student Evaluations of Teaching. Retrieved April 30, 2019, from https://cft.vanderbilt.edu/guides-sub-pages/student-evaluations/

McKeachie, W. J. (1997). Student Ratings: The Validity of Use. American Psychologist. 52(11),1218-25. Available at ISU Parks Library, General Collection: BF1 Am38

Mengel, F., Sauermann, J., & Zolitz, U. (2018). Gender Bias in Teaching Evaluations. Journal of the European Economic Association, 17(2), 535-566 doi:10.1093/jeea/jvx057

Merritt, D. J. (2007). Bias, the Brain, and Student Evaluations of Teaching. SSRN Electronic Journal. doi:10.2139/ssrn.963196

Mitchell, K. M. W., & Martin, J. (2018). Gender Bias in Student Evaluations. PS: Political Science & Politics, 51(03), 648–652. doi:10.1017/s104909651800001x

Nevo, D., McClean, R., & Nevo, S. (2010). Harnessing Information Technology to Improve the Process of Students' Evaluations of Teaching: An Exploration of Students' Critical Success Factors of Online Evaluations. Journal of Information Systems Education, 21(1), 99-109.

Oermann, M., Conklin, J., Rushton, S., & Bush, M. (2018). Student evaluations of teaching (SET): Guidelines for their use. Nursing Forum, 53(3), 280-285.

Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. M. T. (2007). A meta-validation model for assessing the score-validity of student teaching evaluations. Quality & Quantity, 43(2), 197–209. doi:10.1007/s11135-007-9112-4

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.Com. Journal of Diversity in Higher Education, 3(3), 137–152. doi:10.1037/a0019865

Rezaei, M., Ghartappeh, A., Bagher Kajbaf, M., Safari, Y., Mohammadi, M., & Sharafi, K. (2018). Validating "Students' Opinion Questionnaire" and "Student's Evaluation of Educational Quality Questionnaire" in Relation to Teacher Evaluation Using Criterion Method. Educational Research in Medical Sciences, 7(1), Educational Research in Medical Sciences, 01 July 2018, Vol.7(1).

Riniolo, T. C., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006). Hot or Not: Do Professors Perceived as Physically Attractive Receive Higher Student Evaluations? The Journal of General Psychology, 133(1), 19–35. doi:10.3200/genp.133.1.19-35

Risquez, Anglica, Vaughan, Elaine, & Murphy, Maura. (2015). Online Student Evaluations of Teaching: What Are We Sacrificing for the Affordances of Technology? Assessment & Evaluation in Higher Education, 40(1), 120-134.

Ryalls, K., Benton, S., & Li, D. (2016). Response to "Zero correlation between evaluations and learning". IDEA Editorial Note #3. Manhattan, KS: Kansas State University Center for Faculty Evaluation and Development. Retrieved on February 5, 2019 from http://www.ideaedu.org/portals/0/uploads/documents/Response_to_Zero_Correlation_Between_Evaluations_Teaching.pdf

Smith, B. P., & Hawkins, B. (2011). Examining Student Evaluations of Black College Faculty: Does Race Matter? The Journal of Negro Education, 80(2), 149-162.

Standish, T., Joines, J., Young, A., & Gallagher, K. (2018). Improving SET Response Rates: Synchronous Online Administration as a Tool to Improve Evaluation Quality. Research in Higher Education, 59(6), 812-823.

Stanny, C., Arruda, J., Gurung, Regan A. R., & Landrum, R. Eric. (2017). A Comparison of Student Evaluations of Teaching With Online and Paper-Based Administration. Scholarship of Teaching and Learning in Psychology, 3(3), 198-207.

Stark, P. B., & Freishtat. R. (2014). An evaluation of course evaluations. ScienceOpen Research, ScienceOpen Research, 01 September 2014.

Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science* 185(4157), 1124-1131. doi:10.1126/science.185.4157.1124.

University Policy Statement 210.002, Tenure and Promotion Personnel Standards. (March 5, 2019). Retrieved from http://www.fullerton.edu/senate/publications_policies_resolutions/ups.php

University Policy Statement 220.000, Policies, Procedures, and Guidelines for the Administration of Student Opinion Questionnaire (SOQ) Forms. (October 8, 2007). Retrieved from http://www.fullerton.edu/senate/publications_policies_resolutions/ups.php

Uttl, B., White, C.A., & Gonzalez, D.W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related**.** Studies in Educational Evaluation, 54(2017), 22-42 doi:10.1016/j.stueduc.2016.08.007

Whitworth, J. E., Price, B. A., & Randall, C. H. (2002). Factors that affect college of business student opinion of teaching and learning. Journal of Education for Business, 77(5), 282-289.

Wigington, H., Tollefson, N., & Rodriguez, E. (1989). Students' Ratings of Instructors Revisited: Interactions among Class and Instructor Variables. Research in Higher Education, 30(3), 331-344. Retrieved from http://www.jstor.org.lib-proxy.fullerton.edu/stable/40195904

Williams, D. A. (2007). Examining the Relation between Race and Student Evaluations of Faculty Members: A Literature Review. Profession, 2007(1), 168–173. doi:10.1632/prof.2007.2007.1.168

FURTHER READINGS

Abrami, P. C. (1985). Dimensions of effective college instruction. *Review of Higher Education, 8*(3), 211-228.

Abrami, P. C. (1993). Using student rating norm groups for summative evaluation. *Evaluation and Faculty Development, 13*, 5-9.

Adams, Meredith J. D., & Umbach, Paul D. (2012). Nonresponse and Online Student Evaluations of Teaching: Understanding the Influence of Salience, Fatigue, and Academic Environments. Research in Higher Education, 53(5), 576-591.

Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley, L., Mohanty, G., & Spooner, F. (2004). Student Evaluation of College Teaching: A Practice in Search of Principles. *College Teaching*. *52*(4), 134. Retrieved December 10, 2007, from http://www.metapress.com/content/f470q86m45xg1467/?p=46bab8cf74e64505ac b584dc45568855&pi=2

Arreola, R. A. (2000). *Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive*

faculty evaluation system (2nd Ed.). Bolton, MA: Anker Publisher Company.
Available at ISU Parks Library: LB2333 A77 2000

Cannon, R. (2001). Broadening the context for teaching evaluation. In C. Knapper & P.
Cranton (Eds.), *Fresh approaches to the evaluation of teaching* (Theme Issue). *New
Directions for Teaching and Learning, 88*, 87-97. Retrieved December 10, 2007, from
http://www3.interscience.wiley.com/journal/89011422/issue

Cashin, W. E. (1999). Student ratings of teaching: Uses and misuses. In P. Seldin (Ed.),
*Changing practices in evaluating teaching: A practical guide to improved
faculty performance and promotion/tenure decisions* (pp. 25-44). Bolton, MA:
Anker. Available at ISU Parks Library: LB2333 S436x 1999

Cashin, W. E., & Downey, R. G. (1992). Using global student rating items for summative
evaluation. *Journal of Educational Psychology, 84* (4), 563-572. Available at ISU Parks
Library, General Collection: LB1051 A1 J8

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining
faculty effectiveness.* San Francisco: Jossey-Bass Publishers. Available at ISU Parks
Library, General Collection: LB2333 .C456 1993

Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An
assessment of student perception and motivation. *Assessment & Evaluation in Higher
Education, 28*(1), 71-88. Retrieved December 11, 2007, from
http://proxy.lib.iastate.edu:2048/login?url=http://search.epnet.com/login.aspx?authtype
= ip,uid&profile=ehost&defaultdb=afh

Cheung, D. (2000). Evidence of a single second-order factor in student ratings of teaching
effectiveness. *Structural Equation Modeling, 7*(3), 442-460. Retrieved December 11,
2007, from http://proxy.lib.iastate.edu:2048/login?url=http://www.leaonline.com/loi/sem

Coen, P. A. (1983). Comment on a selective review of the validity of student ratings of teaching,
Journal of Higher Education, 54, 448-458. Retrieved December 11, 2007, from
http://links.jstor.org/sici?sici=0022-
1546%28198307%2F08%2954%3A4%3C433%3ATATI%3E2.0.CO%3B2-V

Ewing, Andrew M (2012). Estimating the impact of relative expected grade on
student evaluations of teachers Economics of Education Review 31. 141– 154

Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student
evaluations of your teaching accurately. In K. G. Lewis (Ed.). *Techniques and
strategies for interpreting student evaluation* (Theme Issue). *New Directions for
Teaching and Learning, 87*, 85-100. Retrieved December 11, 2007, from
http://proxy.lib.iastate.edu:2048/login?
url=http://www3.interscience.wiley.com/cgi-bin/jtoc?ID=86011233

Freng, Scott, & Webber, David. (2009). Turning up the Heat on Online Teaching Evaluations: Does "Hotness" Matter? Teaching of Psychology, 36(3), 189-193.

Harrison, P. D., Douglas, D. K., & Burdsal, C. A. (2004). The Relative Merits of Different Types of Overall Evaluations of Teaching Effectiveness. *Research in Higher Education. 45*(3), 311-323. Retrieved December 11, 2007, from http://proxy.lib.iastate.edu:2048/login?url=http://www.springerlink.com/openurl.asp?genre=journal&issn=0361-0365

Hobson, S. M., & Talbot, D. M. (2001). Understanding student evaluations: What all faculty should know. *College Teaching, 49*(1), 26-31. Available at ISU Parks Library, General Collection: LB2300 Im71

Hodges, L. C., & Stanton, K. (2007). Translating Comments on Student Evaluations into the Language of Learning. *Innovative Higher Education. 31*(5), 279-286. Retrieved December 11, 2007, from http://www.metapress.com/content/6m576x4615608441/?p=2eb1fb52456540e8804e82 3 aeaff3037&pi=3

Lewis, K. G. (2001). Making sense of student written comments. In K. G. Lewis (Ed.). *Techniques and strategies for interpreting student evaluation* (Theme Issue). *New Directions for Teaching and Learning, 87*, 25-32. Retrieved December 11, 2007, from http://proxy.lib.iastate.edu:2048/login?url=http:// www3.interscience.wiley.com/cgi-bin/jtoc?ID=86011233

Marincovich, M. (1999). Using student feedback to improve teaching. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 45-69). Bolton, MA: Anker. Available at ISU Parks Library: LB2333 S436x 1999

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research, *International Journal of Educational Research, 11*, 253-388. Retrieved December 11, 2007, from http://www.sciencedirect.com/science/journal/08830355

Marsh, H. W. (1995). Still weighting for the right criteria to validate student evaluations of teaching in the IDEA system. *Journal of Educational Psychology, 87*, 666-679. Available at ISU Parks Library, General Collection: LB1051 A1 J8

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective. *American Psychologist, 52*(11), 1187-1197. Available at ISU Parks Library, General Collection: BF1 Am38

McKeachie, W. J. (1990). Research on college teaching: the historical background. Journal of Educational Psychology, 82(2), 189–200. McKeachie, W. J. (1997). Student ratings: the validity of use. American Psychologist, 52(11), 1218–1225. Miller, J. E., & Seldin,

P. (2014).

Murray, H. G. (1983). Low-inference classroom teaching behaviors and student ratings ofcollege teaching effectiveness. *Journal of Educational Psychology, 75*, 138-194. Available at ISU Parks Library, General Collection: LB1051 A1 J8

Murray, J. P. (2001). Reflecting on student ratings: A personal approach to teaching improvement. *Journal of Faculty Development, 18*(4), 131-136. (Not owned by ISU Parks Library; contact Interlibrary Loan to request article: http://www.lib.iastate.edu/isu- bin/ILL)

Neumann, R. (2000). Communicating student evaluation of teaching results: Rating interpretation guides (RIGs). *Assessment & Evaluation in Higher Education, 25* (2), 121-134. Retrieved December 11, 2007, from
http://proxy.lib.iastate.edu:2048/login?url=http://search.epnet.com/login.aspx?authtype= ip,uid&profile=ehost&defaultdb=afh

Rando, W. L. (2001). Writing teaching assessment questions for precision and reflection. In K. G. Lewis (Ed.). *Techniques and strategies for interpreting student evaluation* (Theme Issue). *New Directions for Teaching and Learning, 87*, 77-83. Retrieved December 11, 2007, from
ttp://proxy.lib.iastate.edu:2048/login?url=http://www3.interscience.wiley.com/cgi-bin/jtoc?ID=86011233

Renaud, R. D., & Murray, H. G. (2005). Factorial Validity of Student Ratings of Instruction. *Research in Higher Education. 46*(8), 929-953. Retrieved December 11, 2007, from ttp://proxy.lib.iastate.edu:2048/login?url=http://www.springerlink.com/openurl.a sp?genre=journal&issn=0361-0365

Rice, R. E., & Stewart, L. P. (2000). Extending the domain of instructional effectiveness assessment in student evaluations of instruction. *Communication Education, 49*, 253-266. Available at ISU Parks Library, General Collection: PN4071 Sp351

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: Love me, love my lectures? *Assessment & Evaluation in Higher Education, 25*(4), 397-405. Retrieved December 11, 2007, from http://proxy.lib.iastate.edu:2048/login?url=http://search.epnet.com/login.aspx?authtype= ip,uid&profile=ehost&defaultdb=afh

Sproule, R. (2000). Student evaluation of teaching: A methodological critique of conventional practices. *Education Policy Analysis Archives, 8*(50). Retrieved December 11, 2007, from http://epaa.asu.edu/epaa/v8n50.html

Svinicki, M. D. (2001). Encouraging your students to give feedback. In K. G. Lewis (Ed.). *Techniques and strategies for interpreting student evaluation* (Theme Issue). *New Directions for Teaching and Learning, 87*, 17-24. Retrieved December 11, 2007, from

http://proxy.lib.iastate.edu:2048/login?url=http://www3.interscience.wiley.com/cgi-bin/jtoc?ID=86011233

Theall, M., & Franklin, J. (2001). Using technology to facilitate evaluation. In C. Knapper & P. Cranton (Eds.), *Fresh approaches to the evaluation of teaching* (Theme Issue). *New Directions for Teaching and Learning, 88*, 41-50. Retrieved December 11, 2007, from http://proxy.lib.iastate.edu:2048/login?url=http://www3.interscience.wiley.com/cgi-bin/jtoc?ID=86011233