REPORT OF THE SPECIAL COMMITTEE ON EVALUATION OF TEACHING

JUNE 7, 2019

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The Special Committee on Evaluation of Teaching (SCET) was charged with assessing practices for the evaluation of teaching at UC Davis, at peer institutions, and within the published literature, in order to suggest immediately actionable and longer-term potential changes to improve the evaluation of teaching at UC Davis. In ten meetings from January through May 2019 and considerable time outside of meetings, SCET collected, discussed and distilled these sources of information into this comprehensive report.

It is important to note that UC Davis is not the only institution of higher education that is engaged in evaluating its practices for the evaluation of teaching at this time. Many other universities currently are making, or recently have made, minor or major changes in their practices, and indeed, the UC has recently struck a system-wide task force on this topic. The fact that the UC and other universities are engaged in this process does not mean that actions could or should not be taken locally at this time. There were several points of convergence between the findings of published research, innovations at peer institutions and best practices within specific departments and schools at UC Davis that highlighted what changes could be made across UC Davis to improve the evaluation of teaching.

Our recommendations for these short-term and longer-term changes are presented in the final two sections of this report. Four of the most important recommendations are:

1. There needs to be a culture change regarding the perceived purpose of the evaluation of teaching at UC Davis from being primarily or exclusively for *summative* purposes of determining merit and promotion. The evaluation of teaching also can and should serve the *formative* purpose of developing instructors' professional skills and competencies as educators and the *pedagogical* purpose of enhancing teaching effectiveness and students' learning outcomes.
2. To be effective, the evaluation of teaching must be informed from multiple sources, including (but not limited to) student evaluations of teaching (SET), peer evaluations of teaching, and self-assessment of teaching through reflective practices.
3. SET should be recognized and framed as students' *experiences* of their courses rather than *evaluations* of instructors, with the explicit purpose of students' ratings and open-ended feedback being to enhance the quality and effectiveness of instruction.
4. Peer evaluations should be conducted as interactive collaborations between the peer evaluator(s) and an instructor, with meetings prior to and following any in-class observation in order to discuss goals, expectations, performance and recommendations.

INTRODUCTION AND OVERVIEW

In January 2019, the Academic Senate (AS) formed the Special Committee on Evaluation of Teaching (SCET), with the following charge:

"The Special Committee on Evaluation of Teaching will be to (1) evaluate the current practices for evaluation of teaching for merits and promotions at UC Davis; (2) consider practices for evaluation of teaching at peer universities (e.g., wording of items used in SET; methods to improve response rates for SET; alternative ways to evaluate teaching outside of SET such as peer review); and (3) consider research literature on best practices for evaluation of teaching in higher education. The final report should recommend possible near-term minor revisions to the current practices as well as potential long-term major revisions."

An important distinction should be noted in the title and charge of SCET from the committee struck in 2010, Special Committee on Student Evaluation of Teaching (SCSET). Specifically, "Student" was not in SCET's title, and the "current practices for evaluation of teaching for merits and promotions at UC Davis" include more than student evaluations. Thus, SCET did not limit its scope to considering students' reports on their in-classroom experiences of courses. Student reports are an important component of the evaluation of teaching, but they are one component and do not in and of themselves constitute a comprehensive or sufficient means of evaluating university teaching. Peer observations, self-reflection and other procedures also inform the evaluation of teaching.

The SCET members also recognized at least three distinguishable yet complementary goals in the evaluation of teaching. The charge from AS focused on what could be regarded as the summative goal: The collection and presentation of information on which to judge an instructor's performance as an educator for the purpose of determining appropriate merit and promotion. The evaluation of teaching also can, and we would argue should, serve a formative goal: Informing an instructor of those aspects of course design and delivery that could be improved in order to advance the instructor's professional development and enhance teaching effectiveness. That attention to teaching effectiveness draws focus to the third and possibly most important goal: Increasing the pedagogical quality of educational experiences for our students such that their learning outcomes are improved.

Paralleling our attention to these three goals was a broadly shared concern for accuracy, fairness and equity in the evaluation of teaching. The common perspective of SCET's members is that there is widespread dissatisfaction with current practices for the evaluation of teaching at UC Davis because they are seen as susceptible to inaccuracy, unfairness and inequity. Presumably that is why SCET was formed by AS, and it is a theme that we heard from our academic colleagues across the university. There are at least two facets to this concern. First, there is a remarkable degree of variation (or to put it another way, a striking lack of consistency) in evaluation practices across the departments, colleges and schools at UC Davis; naturally this raises questions about whether instructors in some programs are being advantaged by local practices while instructors in other programs are being disadvantaged. Second, demographic, identity and personal characteristics of instructors are seen as affecting the evaluation of teaching in ways that advantage some instructors and disadvantage others; more specifically, implicit

gender, racial and ethnic biases may differentially affect the evaluations of our diverse instructors. Our communications with other institutions and our examination of the academic literature indicate that concerns about gender, racial and ethnic biases are widely-held and may be justified. An essential element of developing "best practices" in the evaluation of teaching at UC Davis is to ensure that said practices improve accuracy, fairness and equity; indeed, the summative, formative and pedagogical goals of the evaluation of teaching cannot be achieved without improving accuracy, fairness and equity in the evaluation of teaching.

SCET's work should be positioned within state and national contexts. When SCET began its work in January 2019, we learned that UC Santa Barbara was surveying other universities about their teaching evaluation practices. Other universities across the country had recently changed, or were in the process of examining, their teaching evaluation practices, particularly with respect to students' ratings of courses and instructors (e.g., U Oregon: https://senate.uoregon.edu/tag/teaching-evaluations/). In May 2019, the UC Academic Senate struck the UC Course Evaluation Task Force to discuss issues of reliability, ralidity, and bias in student evaluations. Thus, we are in a period of broad interest in, active examinations of, and dynamic adjustments to teaching evaluations. SCET and the UC Davis AS should be mindful of this context, as we can engage with and learn from these parallel and ongoing efforts.

The next three sections of this report present our considerations of the three sources of data on the evaluation of teaching that we were charged with assessing: Local practices at UC Davis; practices at peer institutions; and published literature on best practices. Based on our integration of the common themes and most striking observations from these sources, this is followed by two sections presenting the conclusions we drew: Our recommendations for minor changes to the UC Davis teaching evaluation practices that potentially could be implemented in the short term; and our identification of more substantive changes to teaching evaluation that could be considered for further study and future implementation over the longer term.

EVALUATION OF TEACHING AT UC DAVIS

The landscape of current approaches to the evaluation of teaching at UC Davis could be described as "highly variable." There is very little consistency or commonality across units. In part, the reflects the reality of the differences in instructional practices and goals across very different disciplines. Why would one expect the evaluation of teaching to look the same in, for example, Art and Art History, Mechanical Engineering, and Nursing? Yet, all instruction involves some common elements, such as selection of course materials for appropriate depth and breadth of content, accuracy of instructor knowledge, teaching effectiveness for student learning and engagement, etc. For evaluation of teaching for summative purposes (merits and promotions) to be fair and equitable across instructors and units, there need to be broadly agreed-upon standards of what should be evaluated and how evaluations should be conducted. Although there has been less emphasis at UC Davis on the formative (enhancing professional development) and pedagogical (improving student learning) purposes of the evaluation of teaching, there also are mechanisms for improving these efforts that could share common elements across units.

In this section, we summarize the current practices for evaluation of teaching at UC Davis and identify some of the "best practices" that could be considered for broader implementation.

Student evaluations of teaching (SET)

Regulation 534 of the Davis Division of the Academic Senate pertains to Course Evaluations and states:

"In every course designated by the Committee on Courses of Instruction, all instructors must implement a course evaluation procedure in such a manner as to afford to each student the ability to evaluate the instructor and the course. Such evaluations shall be made available to the instructor after grades for the course have been submitted. The evaluation procedure shall, to the fullest extent possible, preserve the anonymity of the student and restrict the identification of the course instructor to authorized persons only, including the Committee on Academic Personnel and others involved in the academic personnel process and in the selection of course instructors. (En. 4/17/2012)"

As a result of regulation 534, current teaching evaluation practices at UC Davis are primarily driven by COCI (Committee on Courses of Instruction) policy that mandates registered students be given the opportunity to evaluate courses offered for academic credit. COCI defines the minimum elements of Course Evaluation as the following 2 questions, as written, in addition to an opportunity for comments:

1. Please indicate the overall teaching effectiveness of the instructor.
   (5 = excellent; 4 = very good; 3 = satisfactory; 2 = fair; 1 = poor)
2. Please indicate the overall educational value of the course.
   (5 = excellent; 4 = very good; 3 = satisfactory; 2 = fair; 1 = poor)

However, there are numerous Memoranda of Understanding (MOUs) with COCI that delegate oversight of course contents to a school or college, as well as the course evaluation process to the

respective school or college. As a result of these MOUs, and individual choices in various schools and departments, there is marked heterogeneity across campus in how course and instructor evaluations are executed. Some units ask students the minimum two questions only, however, other units ask as many as twelve questions regarding the course and up to ten questions regarding the individual instructors. In those units that ask more than the required two questions, there is variation in whether those questions appear at the beginning, end, or somewhere in the middle of the set of questions, which could affect how the questions are interpreted and responded to by students. Some units also ask questions about teaching assistants and some ask students to reflect upon their own efforts and accomplishments. Many units provide the opportunity for open comments at the end of the evaluation form, often being prompted by such things as - " Please comment on what worked well for you and what could be changed in the future to better aid learning ", "Please provide feedback to this instructor on their teaching effectiveness – What worked well? What could be improved upon? ", "For any of the above where you scored "Strongly Disagree", please provide comments", "Do you have any additional comments not captured above? ".

There is great variation across campus in the completion rates of SET. Variables that appear to influence completion rates include provision of dedicated time in class for completion, use of paper versus on-line SET forms, and whether inducements or penalties are attached to completion. Unlike some other institutions (Stanford), release of grades is not tied to completion of student evaluations by any unit at UC Davis. The School of Veterinary Medicine has very high rates of completion of student evaluations, as it links completion to expected standards of professionalism that are articulated to students. Evaluations are tracked, and repeated (3 or more) failures to complete evaluations result in visits to the Students Affairs Committee.
Although the use of SET as a feedback tool for instructors' own use (formative purpose) is not controversial, there is widespread instructor concern about the use of SET in the merit and promotion process (summative purpose). These concerns pertain to the potential for a variety of implicit and explicit biases to affect SET, including gender and racial bias, and to long-standing debate about the validity of SET for assessing accuracy and completeness of course content, teaching effectiveness and student learning. In some merit dossiers, SET are the only supporting information provided for teaching competency, despite APM 210 mandating that, regarding teaching, "more than 1 kind of evidence shall accompany each review file". There is marked departmental variation in mean SET ratings. It is unlikely that this reflects differences in the quality of teaching by different disciplines, but more likely differences in the student populations (lower division versus upper division, undergraduate versus graduate student and professional school) or disciplinary norms about the meaning of anchor points on rating scales.

As such, the consensus view of SCET is that it would be more appropriate to refer to SET as students' ratings of their "experiences" rather than as students' evaluations. Further, if SET are to continue being used for formative purposes, student ratings should be used with caution and only in conjunction with other measures of educational effectiveness when evaluating instructors.

Peer evaluations of teaching

Best practice should dictate that instructors receive feedback not just from students but from other experienced instructors who can comment on course content and design, as well as delivery, i.e., peer observation. Peer Observation of Teaching is one tool that provides rich, qualitative evidence for teachers, quite different from closed-ended or open-ended SET. When Peer Observation of Teaching is incorporated into university practice and culture, and is conducted in a mutually respectful and supportive way, it has the potential to facilitate reflective change and growth for teachers (Siddiqui et al., 2007).

Some departments use a form of peer observation whereby one colleague is asked to observe a lesson and to provide a narrative evaluation of an instructor's teaching that is included in the dossier based on that single observation. However, this type of peer observation has little constructive or formative utility, as the instructor does not have an opportunity to discuss elements of the course and its teaching (e.g., course design, required materials, instructional approach, etc.) with the observer. Further, with respect to summative utility, such reports typically lack credibility as they often are written as exclusively and effusively positive (a "crony report") and provide no useful information to FPC or CAP when making merit and promotion decisions.

There are local examples of academic units that have begun to implement more comprehensive peer evaluation procedures that are more likely to be effective for summative, formative and pedagogical purposes. The MCB department in CBS has instituted a more robust peer review process for their faculty, especially those who are coming up for promotion (tenure step mostly). This process involves selection of two colleagues who have some familiarity with the curriculum in the course, a pre-observation meeting with the "candidate", discussing their approaches and course resources, observation of approximately two appropriate lectures, and subsequent synoptic write up that accompanies the dossier.

Some of the professional schools have begun to provide a more holistic review of instruction, with greater emphasis on "enhancing" teaching effectiveness rather than (only) "evaluating" instructors. The School of Veterinary Medicine (SVM) currently runs a program of peer observation and coaching with the aim to provide instructors with formative feedback on their classroom sessions (both large and small group). The goals of the SVM program are twofold.

1. To provide a process to enhance teaching (and document progress)
2. To provide a framework for formative and summative peer discussion and self-assessment of teaching strategies.

SVM uses a Peer Observation instrument developed in conjunction with the Teaching Academy of the Consortium of West Region Colleges of Veterinary Medicine (https://teachingacademy.westregioncvm.org/). The instrument is based on pedagogical best practices and was developed by a working group specifically tasked to do so. The instrument consists of 3 parts (see appendix). The primary goal of the process is formative but the Post-Observation component has a summary paragraph and a categorical overall perception of teaching – Emerging, Evident, Exemplary – that accompanies the dossier. There are two

observers for each “observation,” ideally, one familiar with education best practices and one content expert / content familiar observer. SVM provides (workshop) training for observers. Two observations are recommended prior to tenure, and an additional observation is recommended prior to promotion to Full Professor. The first observation is a "range-finder", with the intent that progress / enhancement is demonstrated (and documented) with subsequent observations.

The School of Medicine (SOM) is currently working on implementing a similar program that uses content experts and educational specialists to work with instructors to improve teaching effectiveness and vertical integration of the curriculum.

The peer observation programs in both the SVM and SOM utilize pre-observation and post-observation meetings with individual instructors, in addition to classroom observation. The pre-observation meeting helps instructors reflect on the learning objectives/outcomes of the session being reviewed. Additionally, the instructor is asked specific questions about the session with regards to use of new pedagogies or materials. It is important for a professional school (where courses tend to be team-taught) to also have the instructor consider how the session fits into the overall course/curriculum. The post-observation meeting asks instructors to reflect on the teaching session(s), consider any challenges they encountered and how they might enhance the session going forward. Both peer observation/review programs aim to help instructors become more reflective instructors and to encourage use of effective (evidence-based) pedagogies. The ultimate goal of these programs is to enhance student learning and outcomes.

Self-assessment of teaching

Currently there is no consistency in expectations for or documentation of self-assessment of teaching across instructors or units. The teaching statement within merit and promotion dossiers is where most instructors might be expected to provide a summary of their practices. Yet, some instructors simply list the courses taught, their mean SET scores, and perhaps the number of trainees mentored. This does not necessarily mean that these instructors have not engaged in self-reflection in their efforts to enhance teaching effectiveness; they simply may not see that as something to be reported. Other instructors provide extensive detailing of their pedagogical philosophies, the professional skills development workshops or exercises that they have undertaken, their methods for mid-quarterly assessments of student learning and teaching effectiveness, and how they have used such activities to enhance the in- and out-of-classroom experiences of their students and mentees. CEE provides numerous online, individualized and group-based resources and opportunities for instructors to strengthen their self-assessment and enhance their teaching competencies. The extent to which instructors engage in such activities at CEE, at professional conferences or in other venues varies considerably within units, but also between units as the local departmental pedagogical culture appears to shape individual instructors’ attitudes and practices for self-assessment.

EVALUATION OF TEACHING AT PEER INSTITUTIONS

UC Davis is not alone in its efforts to revisit and revise practices for evaluation of teaching. We hoped to gain inspiration and feedback from other institutions engaged in this effort. To this end, we gathered information from other institutions in three ways. First, we solicited feedback from 10 peer institutions (letter in Appendix) and received direct responses from two: UC Irvine (UCI), which shared the results of a recent study comparing two SET formats, and Yale, which recently completed a revision to the SET used in Yale College. Second, our own Center for Educational Effectiveness (CEE) and members of our committee provided information about teaching evaluation at the University of Kansas (KU) and Harvard Medical School. Finally, we reviewed public websites from University of Oregon, University of Southern California (USC) and Vanderbilt which describe those institutions' policies or reviews. Other institutions that we contacted did not respond in time to be included in this report. We have reviewed these materials with a focus on student evaluations, peer observation, self-reflection and other holistic approaches to evaluation of teaching.

Student evaluations of teaching (SET): Oregon, UCI, USC, Vanderbilt, Yale

SET involves a series of questions, presented in survey format to students. This is the most widespread tool for evaluating teaching, but as outlined elsewhere in this report, there are a number of concerns about this approach, particularly as the sole form of instructor evaluation. Other institutions are taking numerous and varied steps to improve SET and use student survey results as a source of information about student experience rather than instructor effectiveness.

- **Changes to SET structure or design**
    - Emphasize reflection on learning outcomes by placing open-ended questions about this first, or early on in the evaluation (UCI, Yale).
    - Indeed, UCI compared two SET formats within 35 courses and over 5,000 students. They found that their newer format resulted in slightly lower ratings for instructor behaviors, but that largely, there was little difference in outcomes, despite numerous changes to the design. Instructor satisfaction was slightly higher with the new form, but the bias identified in other studies (men > women, non-STEM > STEM, smaller courses > larger ones) was detectable in both versions.
    - Focus on the course rather than the instructor: at Yale, only 1 out of 10 questions asks about teaching effectiveness, as opposed to characteristics of the course such as workload, intellectual challenge, and organization.
    - Yale did not compare their old and new forms; they redesigned and now use the updated version.
    - USC recommends SET should include elements addressing: a) instructional design, b) instructor characteristics, c) learning experiences, d) assessments and feedback, e) diversity and inclusion practices.

- **Changes to SET presentation to students**
    - Include introduction about unconscious bias before students take survey (Yale).
    - Set aside time in class for SET, including introduction to the importance of the process, as part of a culture of commitment to teaching (Vanderbilt).

    - Increase the frequency of SET by centrally administered midterm student experience surveys (Oregon).
- **Include questions about students or tie SET results to student outcomes**
    - Include questions about students' effort or commitment to the course (UCI).
    - Include questions about the perceived level of challenge associated with the course (Yale, UCI).
    - Link SET results to student outcomes such as a) project samples as part of student portfolios and performance; b) learning outcomes (recommended at USC); or (c) grades (tried by UCI in their study of 35 courses).
- **Changes to use and presentation of SET findings**
    - Change the name or clearly acknowledge that SET describe student experience or satisfaction, rather than evaluation of teacher effectiveness (Oregon, USC).
    - Make SET course ratings more transparent and easily available, which has resulted in more referrals of instructors to their teaching center (Yale).
    - Include a "warning label" and guidance about bias when SET results are presented in the M&P process (Oregon).
    - Stop using SET as the only form of instructor teaching evaluation (Oregon), or implement a multi-modal evaluation requirement for lecturer promotion process (Yale).
- **Improve response rates**
    - Offer some form of incentive to students for completion (UCI).
    - Suggest tying SET to release of course grades (USC).
    - Set aside time to complete SET in class (Vanderbilt).

Peer evaluations of teaching: Harvard, KU, UCI, USC

As noted earlier in this report, peer observation of teaching has been a component of UCD teaching assessment for tenure and promotion, but there is great variation in its implementation across campus. Other institutions have taken the following actions to develop and improve the peer observation process as part of enhancing, and assessing, teaching effectiveness.

- Harvard Medical School has a handbook and rubric for peer-evaluation process, which mirrors one of the recommendations in Section 5 (recommendations for the short-term) about use of a standardized form.
- USC recommends incorporating peer evaluation into *formative* teaching evaluation, echoing recommendations in Section 5 about the value of peer evaluation in a formative approach to improving instructional quality.
- USC recommends incorporating peer review of *instructional design* (see elements of teaching portfolio below) as part of both the formative and summative evaluation of instructors.
- UCI provides resources for best practices in peer evaluation to be used either on a voluntary basis or as part of merit and promotion that includes some of the recommendations outlined in Section 5, including a pre and post observation meeting between the instructor and observer(s) and use of a standardized form. Other recommendations outlined in Section 5 are not included in the UCI recommendations, as they indicate that their entire process should take only 3 to 4 hours.

Self-assessment of teaching: KU, UCI, USC, Vanderbilt

Self-assessment can include reflection statements, teaching portfolios, or midterm feedback from students. The degree to which any of these are mandated by the institution varies.

- USC and Vanderbilt both emphasize development and use of a teaching portfolio which can include a statement of teaching philosophy, sample syllabi, instructional plans, assignments with grading rubrics, sample work from students, and a self-reflection statement.
- KU and Vanderbilt ask for consideration of how the instructor's teaching has changed over time.
- UCI encourages the use of self-reflection statements and a Teaching Practices Inventory as part of the merit and promotion process.
- Oregon has increased the frequency of SET by adding centrally administered midterm student experience surveys whose results are only available to instructors for formative, not summative, purposes (e.g., to be made aware of course elements that are not working out as planned in order to adapt them accordingly in the second half of the course).

Other approaches: beyond SET, peer observation and self-assessment towards a more holistic approach to instruction quality

Several institutions value evaluation of teaching as a component of campus climate and may also use it as a tool to engage the wider community of educational institutions on a national level.

- USC outlines five specific recommendations about promoting a culture of excellence in teaching on their campus, including:
    - systemic review of course evaluation
    - articulating components and levels of what constitutes teaching quality
    - tying teaching evaluation to their campus-wide diversity plan
    - incentivizing professional development to promote teaching excellence
    - use of their equivalent of our Center for Educational Effectiveness as the platform for implementing change
- KU includes 7 elements of evaluation of teaching and provides a rubric for evaluation of each:
    - goals, content and alignment
    - teaching practices
    - achievement of learning outcomes
    - classroom climate and student perceptions
    - reflection and iterative growth
    - mentoring and advising
    - involvement in teaching service, scholarship or community
- KU and Vanderbilt assess instructor contributions to the broader, off campus teaching community.
- KU shares their teaching effectiveness benchmarks with other institutions engaged in similar efforts.

LITERATURE REVIEW

We reviewed the literature to identify best practices in the evaluation of teaching and to outline factors departments should consider when reflecting on teaching quality. We approached this review having already identified the need for a cultural shift that moves the evaluation process away from one in which teaching evaluations are solely used to sum up the delivery of instruction to a focus on improving teaching and learning. As Hattie (2015) explains, this cultural shift should move departments to focus on "seeking evidence to support interpretations about impact, having collective discussions about this impact, what the magnitude of this impact should be, and how pervasive is this impact on the students" (p. 89-90). That is, we view the evaluation process as an opportunity for instructors to reflect on their instruction with the explicit goal of improving student learning—to be effective, teaching evaluation systems must focus on improvement.

Evaluation of teaching is a process separate from the instrumentation used to gather information about instruction. Evaluation should involve many components, with some results shared summatively—attached to a merit or promotion dossier—and other components used formatively—for instructors to use to reflect on (and to improve) their instruction before the quarter has ended. Formative assessments may be developed and administered externally (e.g., classroom observation tools that are scored by colleagues) or involve the instructor's own data collection for self-assessment, to privately reflect on their teaching.

Finally, we present general principles for the evaluation of teaching but do not prescribe a particular approach in recognition of the fact that each unit will need to develop or tailor a method that is appropriate for their particular context. We share recommendations for three assessment approaches that should be used in conjunction with each other below: student evaluations of teaching surveys, classroom observation instruments, and self-assessment.

Student evaluations of teaching (SET)

SET are surveys that ask students about their experiences in a given class and are used to summarize student experiences. They are one of the most widely used and most studied educational measures, yet little consensus exists about their quality. Hattie (2015) and Marsh (2007) report that SET are reliable and valid indicators of instruction. However, others have noted tremendous variability between students in rating the same instructor (Clayson, 2018), variability in the honesty of evaluations because students sometimes think that ratings could affect their own grades (McClain, Gulbis & Hays, 2018), and potential sources of bias associated with the gender, race, ethnicity, or culture of the instructor (Linse, 2017; Macnell, Driscoll, & Hunt, 2014).

Despite extensive research, the extremity of bias is unknown. As Linse (2017) argues, while bias clearly plays a role in the ratings provided by some students, it is less likely that bias is pervasive enough in most situations to cause persistently low ratings across all courses taught in a review period. Linse goes on to note severe methodological flaws in many studies of bias in SET that lead to incorrect conclusions. This variability in the rigor with which studies are conducted is one factor contributing to the lack of clarity about the extent to which bias substantially pollutes

scores (see also McClain, Gulbis & Hays, 2018 who concur with this interpretation of the literature).

Despite these concerns, SET are one important component of a balanced teaching evaluation system. Recommendations from the literature on SET include:

- Consider class context when interpreting ratings. Things like class size, whether the course is for undergraduate or graduate/professional students, and course topic can influence ratings (Linse, 2017).
- Examine the entire collection of SET ratings instead of focusing on ratings from only one class (Linse, 2017).
- Take active steps to increase response rates, such as: providing students with access to the course evaluation ratings, regularly collecting feedback from students in a variety of formats throughout the quarter so that students understand the value of feedback to instructors, verbally explaining the ways in which feedback will be used and its value to instructors, or administering ratings live in class instead of expecting students to complete them outside of class (Linse, 2017).
- Remember that SET provide important information about the student experience, but that students are unqualified to reflect on important aspects of instruction like the accuracy of instructional content. Recent meta analyses have found inconsistent relations between SET ratings and measures of student learning including course grades, which tend to be small or nonsignificant when examined using high quality SET and learning outcomes measures (Clayson, 2009; Uttl, White, & Gonzalez, 2017).

Peer evaluations of teaching

Peer observations of teaching can also yield helpful information about teaching quality, and they have the added benefit of creating an opportunity for colleagues to jointly reflect on potential approaches for improving instruction. Observations can be useful in their own right as a supplement to SET to provide multiple sources of information about teaching. However, peer observations are much more valuable if the observer and instructor being evaluated meet both before and after an observation session to discuss instructional goals, what happened during the lesson, and next steps in instruction. This formative approach is likely to improve instruction, but requires a commitment on the part of instructors to approach observation as more than checking off a box of tasks necessary to complete a dossier.

Recommendations from the literature on observational approaches to evaluate teaching (discussed more fully in Fletcher, 2018) include:

- Hold a pre-observation meeting to discuss instructional goals, and things the instructor would like the observer to focus on to ensure feedback improves instruction.
- Use a standardized form, rooted in pedagogical best practices, to summarize observation findings. This ensures that important facets of an observation are addressed and ensures consistency across instructors — consistency in observation is especially important if they part of the impact merit and promotion process.

- Remember that instructional quality varies from lesson to lesson and class to class. Studies of classroom observation ratings in the K-12 setting have found that at least five observations are needed to gather reliable ratings (Hill, Charalambos, & Krat, 2012; Kane, Kerr, & Pianta, 2014). While it is unlikely that units will conduct multiple observations per instructor, it is important to keep this in mind in drawing conclusions based upon only one observation.
- Hold a post-observation meeting to discuss the lesson. Feedback should be positive and supportive, inform next steps in instruction, and focus on pedagogy. Observers should address a limited number of issues, ones that are most likely to improve student experiences. Providing too much feedback will be overwhelming for the person being observed and thus limit its utility.

Self-assessment of teaching

Self-assessment is essential to becoming an effective instructor. As Hattie (2015) explains "To be successful, university teachers need to think of themselves as evaluators and ask about the merit, worth, and significance of the impact of their interventions— essentially, successful educators actively practice the Scholarship of Teaching and Learning (SoTL)." (p.80). That is, good teaching is based in evaluating how your instruction is going for all students.

Self-assessment can take many forms, but always involves collecting evidence, reflecting on it, and making instructional changes based on the results. That is, it involves gathering feedback about how instruction is going with the explicit intent of improving instruction. From a promotion, tenure, and merit standpoint, it is essential that instructors write about their self-assessment practices in their teaching statement and that they emphasize what they learned about their teaching and how their instruction changed as a result. Summaries of self-assessment findings are secondary to how instructors respond to results.

Recommendations from the literature on self-assessment include:

- Collect self-assessment evidence during instruction to allow for instructional adjustments. As Hattie (2015) explains, "The most critical mind frame is 'know thy impact'--when an academic walks into a teaching situation their fundamental question needs to be 'how will I know my impact today.'" (p. 89)
- Ensure the anonymity of students in the self-assessment process. As McClain, Gulbis and Hays (2018) point out "…some student responses vary according to which format they think best protects their identity." (p. 381)
- Make sure to include opportunities to share open-ended feedback so key concerns (or compliments) are captured. One approach, recommended by Boston University's Center on Teaching and Learning is to ask three questions: (a) "What should I start doing?" (b)"What should I stop doing?" and "What should I continue doing?" (http://www.bu.edu/ctl/teaching-resources/start-stop-continue/) Other approaches to collecting self-assessment data can be found here: http://www.crlt.umich.edu/gsis/p9_1
- Analyze results and decide how to act upon the results. Share both results and action steps with students during the class meeting immediately following data collection. See

http://www.bu.edu/ctl/teaching-resources/start-stop-continue/ for an example of how to do this.

- Consider the use of teaching portfolios (dossiers) as they are one promising method for presenting self-assessment evidence. Developing a portfolio is an extensive process which involves:
    - Creating a repository to collect work samples. At this stage, units must make decisions about what kinds of work samples to collect and what competencies to address in the portfolio.
    - Developing a procedure to provide feedback on the work samples so that they can be revised, or new work samples added, before the portfolio is finalized.
    - Curating the portfolio. Instructors use feedback to select a portion of work samples for inclusion in the final portfolio (dossier). They also write a reflection that discusses how the work samples provide evidence of meeting the competencies being evaluated (Clarke & Boud, 2016; for details of how to do this and example forms see: https://cft.vanderbilt.edu/guides-sub-pages/teaching-portfolios/)

Conclusions

As Benton and Young (2018) explain:

> *Units that take a balanced approach recognize the challenges in evaluating teaching effectiveness. The accumulated evidence must come from multiple sources and include materials such as descriptions of teaching activities, modifications made to courses, adoption of new teaching strategies, participation in professional-development activities, and contributions made to better the unit's overall instruction. Multiple measures increase the likelihood that the evaluation will encompass all dimensions of teaching, including course design, course delivery, course assessments, instructor availability, and course management." (p.8)*

Treating student evaluations of teaching, peer observation, and self-assessment as three legs of a balanced assessment system is a good first step in creating an environment that encourages support and growth in order to improve teaching practice. It would also improve our conceptualizing of what teaching involves and would more comprehensively capture instructional endeavors.

RECOMMENDATIONS THAT MAY BE ACTIONABLE IN THE SHORT TERM

General

1. Clearly and consistently document in all communications that teaching evaluation requires information from multiple sources, including (but not limited to) SET, peer observation and self-reflection, and that no one source of information should be considered sufficient for the summative purpose of evaluating educator performance.
2. Begin working on culture change from teaching evaluation as *summative* for the purposes of merit and promotion to *also* being *formative* for instructors' development of professional skills and *pedagogical* for enhancing teaching effectiveness and students' learning outcomes. This can begin with simple language changes in the framing and presentation of the various components of teaching evaluation, e.g., wherever possible, replace 'evaluate' with 'enhance' or 'benefit' when describing the purpose (e.g., peer observation for the purpose of *enhancing* teaching performance).
3. Require departments to vote separately on each aspect of evaluation (teaching, research, service) as its own content area prior to voting on recommended merit step for the candidate. Some departments already do this, but not all.
4. Drawing on the expertise of Affirmative Action and Diversity Committee, STEAD, Vice-Chancellor of Diversity and Inclusion and other sources of expertise on diversity and bias, include standard language preceding and introducing teaching dossiers for FPC and CAP members conducting merit and promotion reviews.

Student evaluations of teaching (SET)

1. Rather than "evaluation", characterize SET as "Student ***experience*** of teaching", "Student ***experience*** of the course" or similar (Given AS policy, this may be a long-term recommendation). Assuming the language of the regulation cannot be changed easily, include standard language introducing and framing the online and paper SET forms as being "for the goals of sharing your experience of the course in order to enhance the quality and effectiveness of instruction."
2. Recognizing that we are in a period of local (UCD), system-wide (UC) and nation-wide examinations of the use of SET for summative purposes, AS may want to consider including a letter to campus FPCs and CAP explaining the issues relating to reliability and validity of SET for all merit and promotion decisions until there is greater consensus on their appropriate design, administration and interpretation.
3. Encourage all instructors to provide in-class time for students to complete SET, even if doing so online, in order to increase participation. Encourage instructors to introduce SET by discussing the value of SET in helping them to improve future versions of the class (formative and pedagogical goals).
4. Encourage instructors and departments to use more than the minimum two SET questions required by SET, and to precede those two globally evaluative questions with more concrete questions about specific aspects of the course (e.g., clarity of presented materials, effectiveness of text or other instructional materials, etc.), instructor (e.g., organization, timeliness, responsiveness, etc.), and student (e.g., attendance, completion of assignments, participation, expected grade). Which specific questions are appropriate

are likely to vary by course and discipline, but knowing which aspects of the course or instruction were less versus more well-received by students is important for the formative goal, and to scaffold the students' understanding and consideration of the two globally evaluative questions.

5. Drawing on the expertise of Affirmative Action and Diversity Committee, STEAD, Vice-Chancellor of Diversity and Inclusion and other sources of expertise on diversity and bias, include standard language introducing online and paper SET forms reminding students of the possibilities of implicit biases when completing SET.
6. Currently the SET scantron forms include demographic information but the online forms do not. This inconsistency should be resolved in one of two ways: Either remove from the scantron forms or add to the online forms. A primary argument for removal would be eliminating the possibility of identifying students in small classes. A primary argument for inclusion would be to increase the potential for instructors and the Center for Educational Effectiveness (CEE) to do research on SET responses in order to identify constituents for whom instruction is working less well and adapt the course correspondingly (formative and pedagogical goals). If demographic is included, currently Gender only has binary "female/male" options; add a non-binary option (at least in class sizes where this would be unlikely to identify individuals).

Peer evaluations of teaching

1. Encourage peer observers to meet with instructor prior to observation to establish expectations for in-class experience, and again after in-class observation to discuss and provide feedback.
2. Encourage peer observers to examine course portfolio (syllabus, materials, Canvas site, etc.) comprehensively prior to in-class observation and as part of overall evaluation.
3. Drawing on CEE resources, academic units should provide a rubric or template of content areas, competencies and pedagogical goals to guide in-class observations and peer evaluations and to increase consistency of peer evaluations across instructors.
4. Re-brand peer evaluations as a "coaching model" of enhancing teaching effectiveness (see above note about culture change and language).
5. Drawing on the expertise of Affirmative Action and Diversity Committee, STEAD, Vice-Chancellor of Diversity and Inclusion and other sources of expertise on diversity and bias, include standard language reminding peer evaluators of the possibilities of implicit biases when completing observations and examinations of course materials.
6. Treat peer evaluation/observation as a facet of department/university service.

Self-assessment of teaching

1. Encourage instructors to adopt the practice of making teaching portfolios/dossiers to more comprehensively document their course design, materials, syllabi, etc.
2. Remind instructors of CEE's available services for teaching reflective practices for effective instruction, for consulting on course design and delivery, etc.
3. Encourage instructors to use mid-quarter SET (ratings or open-ended questions) to learn about students' experiences of the course in time to make instructional adjustments to that course.

4. Remind instructors to document and report efforts made to enhance course delivery and improve their own teaching effectiveness, including any professional-development activities, use of new teaching methods or technologies, etc.

RECOMMENDATIONS FOR ACTIONS THAT MAY BE TAKEN OVER THE LONGER TERM

General

1. Continue working on culture change from teaching evaluation as *summative* for the purposes of merit and promotion to *also* being *formative* for instructors' development of professional skills and *pedagogical* for enhancing teaching effectiveness and students' learning outcomes.
2. Encourage academic units to develop their local standards, expectancies and guidelines for effective teaching evaluation within their unit, for example, identifying the specific questions that are most relevant to include in SET for that discipline, the instructor competencies and course features that should be the focus of peer evaluators, and the particular contents of teaching portfolios and statements that instructors should prepare.
3. Adapt MIV to accommodate more comprehensive teaching portfolios for instructors to document multiple aspects of their courses (syllabi, instructional materials, links to Canvas, etc.).

Student evaluations of teaching (SET)

1. Change the wording of the AS policies and regulations such that, rather than "evaluation", SET are characterized as "Student ***experience*** of teaching", "Student ***experience*** of the course" or similar.
2. No source of information about teaching is without potential bias, but concern was strongest about the validity and usefulness of SET for summative purposes. Some SCET members advocated for AS removing SET from merit and promotion decisions by FPC and CAP (e.g., Oregon) and basing the evaluation of teaching for summative purposes on other sources of information (peer evaluation, self-assessment). Other members of SCET supported retaining the use of SET, with carefully implemented improvements, as one component of a broader assessment of teaching. This topic warrants further study and consultation with vested university constituents prior to deciding on a course of action.
3. Include information on the class mean grade and grade distribution, and departmental norms and ranges for SET and grades, to accompany candidate's SET when FPC and CAP engage in teaching evaluations.
4. Transfer responsibility for running online SET from CAES to Center for Educational Effectiveness (CEE), which already runs scantron SET, in order to increase potential for flexibility and research. Develop mechanisms for CEE to be able to match de-identified (name and SID removed) information on student performance/grades and characteristics (e.g., major, transfer student status, etc.) with SET responses. This would support CEE's ability to do research on relations between SET and learning outcomes, and on instructional effectiveness across different campus constituencies. This will involve defining a minimum class enrollment size for matching to reduce risks of identifying individuals through responses.
   There are perceptions of potentially systematic differences across fields and disciplines with respect to students' norms for completing SET (i.e., students in some units give

lower ratings than students in other units that are not due to differences in instructional quality). This may be an appropriate question for CEE to research.

5. Have CEE partner with Affirmative Action and Diversity Committee, STEAD, Vice-Chancellor of Diversity and Inclusion and other sources of expertise on diversity and bias, to examine the specific wording of items and content areas in SET
6. If actions to improve the validity of SET are put in place, consider the development of a platform for sharing SET for courses (by course, not by instructor) with students (e.g., Yale). Currently students share their opinions about courses informally and on platforms of dubious quality and accuracy (e.g. RateMyProfessor). Sharing accurate SET data may help to combat such inaccurate information.
7. Implement mid-quarter SET for instructors to learn about students' experiences of the course in time to make adjustments. These should be formative and for instructors' ability to enhance the course, not summative for the purposes of merit/promotion evaluations, and hence shared only with the instructors (and, possibly, the students).
8. Avoid individual carrot-or stick approaches (e.g., student gets an extra point for completing SET, or student does not get grade released until SET completed). Study whether establishing collective reward procedures (e.g., all students get extra point if $\geq$ 85% of class completes SET) may be appropriate as part of an effort to advance the cultural shift toward effective and inclusive teaching evaluation.
9. Outliers may introduce statistical bias in the interpretation of SET. Consider implementing methods to attenuate this, for example, removing the extreme tails of the distributions (trim the top and bottom of the student responses symmetrically, such that a percentage of the lowest scores and the same percentage of the highest scores are excluded from the average score).

Peer evaluations of teaching

1. Establish protocol for observer(s) to meet with instructor prior to observation to define expectations and instructor's goals for the class, and to meet with instructor after the observation to discuss the class experience.
2. Develop training and support structure for conducting more effective peer observations. This could include an in-person workshop or an online training module for peer observers to complete prior to meeting with instructors, guidelines about expectations for effective observation and feedback, a rubric of rating scales and open-ended prompts on specific aspects of class instruction for observers to complete as part of their evaluation, criteria for evaluating course syllabi and other materials, etc.
3. Having more than one observer and/or observing more than one in-class instruction should become the norm in order to have more comprehensive and accurate, and likely less individually biased, feedback and evaluation by peer observers.
4. Consider observer teams including both a within-department or within-discipline member (for content) and a member from outside the department (for greater objectivity).
5. Improving the effectiveness of peer evaluations for formative, pedagogical and summative purposes is likely to increase the time involved in conducting peer evaluations. AS needs to conduct a cost/benefit analysis of this, and determine appropriate methods of compensating peer observers for their time (e.g., establish this as formal university service).

Self-assessment of teaching

1. Develop a rubric for guiding self-evaluation; for example, a template that parallels the content areas of the rubric for peer observations may be effective.
2. Provide examples of self-assessment strategies that instructors can use to evaluate their own performance, including use of mid-quarter SET to obtain student feedback on experience of course and instruction in time to make adjustments.
3. Create a platform with examples of effective teaching portfolios/dossiers and teaching statements with guidelines for what should/could be included in these, including description of self-assessment procedures in the teaching assessment (Vanderbilt, USC).

REFERENCES

Benton, S. L., & Young, S. (2018). Best Practices in the Evaluation of Teaching. *IDEA Paper*, *69*.

Clarke, J. L., & Boud, D. (2018). Refocusing portfolio assessment: Curating for feedback and portrayal. *Innovations in Education and Teaching International*, *55*(4), 479-486.

Clayson, D. E. (2018). Student evaluation of teaching and matters of reliability. *Assessment & Evaluation in Higher Education*, *43*(4), 666-681.

Fletcher, J. A. (2018). Peer observation of teaching: a practical tool in higher education. *The Journal of Faculty Development*, *32*(1), 51-64.

Hattie, J. (2015). The applicability of Visible Learning to higher education. *Scholarship of Teaching and Learning in Psychology*, *1*(1), 79.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64.

Kane, T., Kerr, K., & Pianta, R. (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. John Wiley & Sons.

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

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383). Springer, Dordrecht.

McClain, L., Gulbis, A., & Hays, D. (2018). Honesty on student evaluations of teaching: effectiveness, purpose, and timing matter!. *Assessment & Evaluation in Higher Education*, *43*(3), 369-385.

Siddiqui, Jonas-Dwyer and Carr. Medical Teacher 2007; 29: 297-300.

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,* 22-42.