Report of the Ad Hoc Working Group on Course Evaluations

March 25, 2019

Members: David Schwartz (co-chair), Kathie Hendley (co-chair), Peggy Hacker, Cecilia Klingele, Mitch, Carrie Sperling, Trina Tinglum, Jenny Zook

I. Executive Summary

The Law School's current Student Evaluation of Teaching (SET) form (officially entitled "course evaluation" form) has significant deficiencies in its reliability, validity, and administration. Even the best-designed and -administered SET forms raise serious questions about sex and race bias. But the UW Law SET form, although well-intentioned, is not well-designed, and its deficiencies increase the risks of sex and race bias in addition to its general reliability and validity problems. These deficiencies in the UW Law SET form render its current use as a significant metric in faculty tenure, promotion, and performance-evaluation decisions highly problematic, raising concerns about the form's compliance with federal antidiscrimination law.

For the reasons set forth in this report, the Working Group makes the following recommendations:

1. **Reporting of SET scores.** The Law School should indefinitely suspend reporting numerical SET scores to faculty and students, effective Spring semester 2019 until the implementation of a valid and reliable long-term SET protocol (see recommendation 4). SET score reports currently made available to students should be removed from the Law School web site. The Law School should continue to collect the numerical data for purposes set out in recommendation number 2.

2. **Use of SET data.** On an interim basis, pending the implementation of a valid and reliable SET procedure (see recommendation 4), the use of numerical SET scores for tenure, promotion, and performance evaluation decisions should be suspended, effective immediately. Qualitative comments should be distributed to instructors for their use in self-assessment of teaching. Qualitative and numerical data from the current SET form should not be used as conclusive evidence of teaching quality, but may be considered as an indicator of need for further supervisorial inquiry into classroom problems. Qualitative and numerical data from the current SET form may continue to be used for determination of adjunct professor renewal. All uses of any data from current SET forms should be with due regard for their reliability, validity, and bias problems.

3. **Revision of applicable Tenure Rules.** The Law School's Tenure Rules should be promptly amended to conform to recommendation 2.

4. **Standing Committee to revise SET protocol.** A standing faculty committee on SETs should be convened. As a first order of business, it should create a new interim evaluation form within the next year. The form should emphasize student self-assessment and objective facts about teaching practices rather than subjective teacher evaluation. The standing committee on teaching evaluation should thereafter study and propose a teaching evaluation system, and a reliable and valid student feedback protocol for long-term use, based on consultation with experts in the field. The standing committee should expeditiously determine whether and how SET data should be used in nominating law school faculty for university teaching awards.

II. The UW Law "~~Course~~ Teaching" Evaluation Form

The UW Law School SET form is designated a "Course Evaluation," and this Working Group is similarly named, but this is a misnomer. The questions on the SET form are all aimed at evaluations of the **instructor**, rather than the substance of the **course**, so they are more aptly described as Student Evaluations of *Teaching*.

In discussing the UW Law SET form, this report will follow the technical distinction drawn in the scholarly research on SETs between **reliability** and **validity.** A **reliable** test or survey instrument provides consistent measures from one iteration to the next. A **valid** instrument accurately measures the target phenomenon. An instrument can be reliable but not valid, by producing consistent results but failing to measure the target phenomenon. For example, a SET question asking students to "agree," "strongly agree," or "disagree," etc. with the statement "The professor provided delicious treats on the last day of class" may be reliable but not a valid measure of teaching effectiveness. Conversely, an instrument can measure the target phenomenon but produce inconsistent results.

There is strong reason to believe that the current SET form in use at UW law school is both invalid and unreliable.

A. Analysis of the Form

The UW Law "Course Evaluation form" (hereinafter "SET form" or "form") consists of five questions asking students to rate their professor or instructor on a five-point likert scale.[1] Students are also invited (but not required) to provide qualitative comments. A substantially similar 5-question form has been continuously in use at the Law School since 1995, with only slight changes to question wording. The most recent change occurred in 2002, when question 5 was re-worded.[2]

The primary purpose of the forms, judging by how they are used, is to measure teaching effectiveness, or "excellence" (in the tenure standard language). But in fact, the forms are a mere intuitive heuristic for teaching evaluation purposes. As Nobel laureate behavioral psychologist Daniel Kahneman explains, "This is the essence of intuitive heuristics: when faced with a difficult question, we often answer an easier one instead, usually without noticing the substitution."[3] More colorfully, economist Darrel Huff made the same point in his satirical

---

[1] The questions are:

Q1. The professor's knowledge of the subject matter of the course.

Q2. The professor's preparation and organization of the class, including organization of the entire course and preparation for each class.

Q3. The professor's personal receptiveness to students, including receptiveness to consultation outside of class.

Q4. The professor's success in getting you to think in greater depth about the topics discussed in the class.

Q5. The overall quality of the professor's teaching in this course.

[2] See Appendix 1. Prior to 1995, the Law School used a SET form with 11 likert questions and 4 open-ended questions.

[3] Daniel Kahneman, Thinking Fast and Slow 12 (2011).

classic, *How to Lie With Statistics*: "If you can't prove what you want to prove, demonstrate something else and pretend that they are the same thing. In the daze that follows the collision of statistics with the human mind, hardly anybody will notice the difference."[4]

The question "what is effective teaching" is extremely difficult, and one that we have not as a faculty answered. The Law School form switches out this question and substitutes an easier one: "how do students feel about the professor's teaching?" The latter question is much easier because the student response is (treated as) self-validating: all students are entitled to their own opinions. The form does not even require an explanation of the opinion, leaving the qualitative exposition optional and providing almost no structured guidance as to what the faculty, as experienced experts, believes good law teaching entails.

The form suggests that a good teacher is "knowledgeable" (Q1), "organized" (Q2), "accessible" (Q3), and able to "get[] [students] to think in greater depth" about the subject (Q4). Even if these four things were somehow a thorough list of what makes a good teacher, the "bottom line" question (Q5) about "overall teaching quality" is presented as another factor rather than as a weighted average of the first four factors. And even if responding students took it upon themselves to treat Q5 as a weighted average of the previous four questions (which is doubtful), the weights are unspecified: is "accessibility" less, equally, or more important than "knowledge"? Every member of the faculty can find examples in her own files of course evaluation forms in which the Q5 number is less than any reasonable weighted average of the previous questions, suggesting that students feel there is a significant factor in "overall teaching quality" that is not captured in Q1-Q4. (The student interface for the SET form is shown in Appendix 2.)

More troubling still, the purported teaching criteria are vague and subjective. They are in fact conclusions based on a black box of more granular but unspecified elements. Are students well positioned to judge an instructor's "knowledge" of the subject, and if so, how do they make that assessment? Does "organization" mean avoiding meandering class discussions, or telling students "what will be on the exam," or something else? Does "personal receptiveness and accessibility" mean responding rapidly to a student's emails, or being available to talk when a student happens by the professor's office, or being "friendly and approachable"?[5] Is it the professor's responsibility to "get" students to "think in greater depth" about the subject? "Greater" than what? Whether or not students are sound judges of professional standards of teaching effectiveness, the form leaves us guessing about whether their responses to the form's conclusory questions are based on valid granular criteria. The SET form's vagueness, subjectivity and lack of structured guidance allow more room for biased responses.

It should be noted that the general Law School SET form is used for all *non-clinical* courses. The clinics employ different course evaluation protocols and forms, one of which is attached as Appendix 7. The Committee has not examined these protocols or forms.

[4] Darrell Huff, How to Lie With Statistics 74 (1954)

[5] Questions 2 and 3 are also ambiguously worded. Are the teaching practices identified after the word "including" meant to suggest the sole factor of "organization" and "accessibility," or a dominant factor, or just one factor among many?

B. Purposes of the Form

Because a test or survey instrument like a course evaluation cannot be validated without a clear specification of its objective or purpose (as discussed further below), it is crucial for the faculty to articulate an agreed-upon purpose for any course evaluation form. While the primary purpose of the Law School form has been to evaluate teaching quality, a form might be justified for other uses. These include (1) measuring **student learning** as opposed to teaching excellence – the two are related, of course, but different; (2) reporting **objective facts** about teaching methods and practices rather than subjective evaluative conclusions; (3) assisting students in their **course selections**; (4) reporting **student satisfaction**; (5) giving students a feeling of "**voice**" in the institution, by viewing course evaluations as a sort of "suggestion box."

1. Teaching and learning feedback. Institutions that have conformed their student-feedback practices to research on teaching evaluations have generally shifted emphasis away from subjective evaluations of teaching toward self-assessment of student learning and the reporting of objective facts about teaching practices. Other than the instructor, the students are in the best position to observe what goes on in a course, both inside and outside the classroom. And while self-reporting is imperfect, an individual student may be best positioned to assess her own learning.[6] Therefore, items 1 and 2 are promising areas for student feedback questionnaires. The current UW form is poorly designed for these purposes, however. Other than Q4, which vaguely gestures at student learning, the form is oriented entirely toward subjective evaluation of teaching, and the questions lack the specificity to generate data about teaching practices, relying entirely on haphazard student qualitative comments for that purpose.

2. Student course selection. A secondary purpose for course evaluations at UW Law School, judging by the way they are administered, is to assist students in selecting courses. Teaching evaluation summaries, showing class averages for the Q1-Q5 likert numbers, are posted for tenured faculty, academic staff (other than legal writing instructors) and adjunct faculty, but not for untenured faculty or legal writing instructors. This practice raises two questions: (1) Are the evaluations summaries valid and reliable information for this purpose? (2) Is this information useful to and wanted by students?

The reliability and validity problems with the form are discussed below. Those aside, one has to ask why the forms are withheld from some categories of instructor. If the information is deemed unfair or unduly harmful for untenured faculty and LRW instructors, what makes it fair and not unduly harmful for tenured faculty?[7] Moreover, might there be a risk of an echo chamber or anchoring effect, in which a professor's past evaluation numbers affect current ones?

---

[6] Philip B. Stark and Richard Freishtat, "An Evaluation of Course Evaluations," ScienceOpen.com (2014) https://www.scienceopen.com/document?vid=42e6aae5-246b-4900-8015-dc99b467b6e4; Betsy Barre, "Student Ratings of Instruction: A Literature Review," Rice University Center for Teaching Excellence (podcast 2015) (hereinafter "Rice Podcast").

[7] The rationale for non-posting of LRW SET scores is said to be that LRW is a required course and thus not relevant for student course selection. But this explanation does not comport with Law School practice. No SET scores of any kind are posted for LRW faculty, even those who teach electives, like Advanced LRW, Legal Correspondence, Trial Advocacy, and upper level legal research courses. In contrast, tenured faculty evaluations for non-elective fall 1L courses (Contracts, Crim, Civ Pro) are posted. Nor does the practice accurately reflect a belief that tenure equals thick skin, because non-tenured, non-LRW instructional staff also have their SET scores posted, including SET scores for required 1L courses.

Given these drawbacks, it is worth asking whether there is a significant student demand for this information. For example, an informal survey conducted by the associate dean at Northwestern Law School suggested that students did not find likert number averages helpful to their course selection decisions, but instead wanted to know specifics about teaching practices: teaching style, how much reading was assigned, and whether or not the professor cold-called on students.[8]

3. Student satisfaction and voice. Student evaluations may be useful for purposes other than measuring teaching quality. Student satisfaction may be an important datum. Giving students a feeling of "voice" may also be a worthwhile goal. But if these are the primary purposes, the survey questions should be more tailored to those objects, and the form should not necessarily be used as a determinant of teaching quality in employment-related decisions.

C. Administration of the Form

The Law School administers its SET forms through a multistep process. The process begins with a reminder email from Dean's office toward the end of the semester, asking faculty to set aside class time for students to fill out the forms on line. Students are sent a sequence of notifications that they may complete the online SET form at any time during the last two weeks of the semester. The process goes on to include compilation of numerical averages and distribution of the completed forms to individual faculty, posting the numerical averages on Law School web pages for the informational benefit of students, and interpreting the forms and employing them in various faculty performance reviews, such as pre-tenure review meetings and tenure dossiers. This administrative process is carried out each semester in good faith, with the best of intentions and commendable effort. Without in any way meaning to diminish these intentions and efforts, the unfortunate fact is that the Law School administers the evaluation forms in a manner that greatly exacerbates the reliability problems inherent in the form itself.

1. Student completion of the form. The SET form is presented to students on a single web page with a small comment box followed by the five questions with "click" (or "radio") buttons for the 1-to-5 rating. (See Appendix 2.) Despite a printed suggestion that "thoughtful comments … often provide helpful information," the form is designed to be filled out fast rather than to promote serious reflection about the course and teaching quality issues. Again, the qualitative comment section is optional. Students choosing to forego it can complete the form in 10 seconds. An informal survey by Dean Kelly suggests that about 1/3 of student SET responders decline to make qualitative comments and that overall, only about 50% of enrolled students offer qualitative comments. This low response rate affects reliability, as discussed below.

2. Interpretation problems. One of the most glaring and troubling aspects of the SET process is the interpretation of the numerical SET scores. On its face, the form advises students to interpret the likert scale as a criterion-based norm, rather than a comparative distributional "grading" curve. The SET form states that 5 = "excellent," 4= "very good," 3= "good," 2= "fair" and 1= "poor." As a first step in the interpretive process, however, the Law School disregards and distorts these linguistic signifiers by transforming this scale into a purely comparative one. Law School administrators and faculty interpret a score of 4 as not "very good" (as the form states), but below average. A score of 3 is not "good," but substandard or "needs improvement."

[8] Interview with Assoc. Dean Sarah Lawsky.

Pre-tenure faculty with class averages between, say, 3.3 and 3.5 – between "good" and "very good" on the form's prima facie language – are counseled of their need to improve. As a practical matter, a score of 2 is not "fair," but poor. It is possible that this linguistic disconnect is due at least in part to the inflated language of the tenure standard, which demands "excellence" in teaching as a general rule. Yet, ironically, the Law School routinely reports that faculty members have met this "excellence" standard on the basis of course evaluations in the mid-fours, corresponding to averages less than "excellent" but better than "very good" – perhaps "very, very good." The particular linguistic disconnect between the SET form and the tenure standards is beyond the scope of this report. But in terms of the practical interpretation given by the Law School to the numerical averages, the numbers might more accurately be labelled: 5=outstanding, 4=almost excellent, and yet below average, 3=needs improvement, 2=poor, 1=disastrously poor.

What converts the disregarded prima facie descriptors ("excellent, very good, etc.") into practical interpretations is the relationship of an instructor's class average to the Law School "mean." The mean for each of the five questions is calculated by averaging the score of each individual SET form submitted by each student for each course. The instructor's classwide mean for a particular course is then compared to the law school mean. The Law School mean for Q5 has been consistently over 4 for at least the past 20 years, ranging from a low of 4.17 to a high of 4.55. (Intriguingly, the mean has trended steadily upward in this time frame.) An instructor who receives a class average of 4 might justifiably claim to have been a "very good" teacher of that particular course, but the Law School's practice is to view that performance as "below the mean," and therefore "below average." This comparative interpretation of the likert averages is problematic, in at least three respects.

First, it creates a norm that is at odds with the prima facie linguistic meaning of the form, and with the tenure standard. The tenure standard of "excellence" is treated, not as a comparative measure, but a norm- or criterion-based one. Criterion-based assessments try to rate every candidate by an objective measure without regard to distribution of outcomes, in contrast to a grading curve. If teaching excellence required attainment of student ratings "above the mean," then about half of tenure candidates should fail to meet the standard. But the tenure standard does not operate this way, either in theory or practice: the Law School and University take the position that every untenured faculty member could in theory meet the teaching standard for tenure. The evaluation form itself is presented as criterion-based, rather than comparative. It suggests that instructors on the down slope of the distribution curve of student ratings can nevertheless be "very good." Indeed, Q5 on the form was changed in 2002 precisely to eliminate comparative ratings. The 1995-2001 version of Q5 expressly asked students to rate the professor "compared with all the instructors I have had in law school."[9] The intent of this change was presumably to embrace a criterion-based rating. Ironically, the fact that students generate a faculty-wide mean well above 3, the midpoint of the scale, suggests that they are reading the evaluation form correctly as criterion-based, rather than comparative. It is the Law School that errs by then transforming the results as comparisons around the mean.

Second, the comparisons made by Law School evaluators are very seat-of-the-pants and impressionistic, while the use of numbers *reported out to two decimal places* creates a false patina of precision. Law school interpreters of the form readily assume that a class average of,

[9] See 1995 form, Appendix 1.

say, 4.29 is "significantly below the mean" of 4.50, yet there is no evidence that such differences are in fact significant. Three disgruntled students out of a class of 50 could easily account for this difference. The Law School does not report or calculate the standard deviation of these averages or otherwise make an effort to validate its interpretations of fractional differences in classwide average scores. Even educational experts who believe that SETs are valid and reliable assert that "faculty with most of their ratings distributed across scores of 3.5–5 on a 5-point scale … are doing well," "even if a few students rate the faculty member at the low end of the scale." Only faculty whose median (not mean) scores fall "below *the midpoint of the scale* likely have an instructional issue" needing attention.[10] The midpoint of the Law School scale is 3, not the faculty mean of 4.5.

Third, and perhaps most glaring of all, the Law School draws comparisons between individual professors' class means and a single law school mean without any effort to control for factors unrelated to teaching quality that may have a significant impact on evaluation numbers. Wholly aside from problems of race and gender bias (about which more later), the Law School erroneously treats classwide likert averages as though they are fully commensurable: an average of 4.9 in a ten-student upper level seminar is deemed comparable to an average of 3.9 in an 80-student first semester 1L course. But research in the field has shown that smaller classes tend to generate more positive evaluation numbers than larger ones.[11] It is also probably the case that elective classes, and upper level classes, will be rated higher, all other things being equal, than required, or first-year classes. [12] Research has also shown that students' individual rating scales change as they become acculturated to the school and gain a wider basis of comparison of instructors so that, for example, freshman and 1L evaluations are not commensurable with upper level student evaluations. Accordingly, researchers assert that where SET scores are used, they must be sorted to account for these differences.[13]

The faculty-wide means for Q5 have steadily increased over the past 20 years, from a low of 4.17 to a high of 4.55. (See Appendix 3.) This is likely further evidence that numerical set scores are statistically noisy. Interestingly, this curve roughly tracks law school enrollments, with lower enrollments corresponding to higher faculty-wide means. [14] This trend is consistent with (a) the research suggesting that non-responding students would probably give lower numerical ratings (see next section) and (b) the contention that Q5 means are influenced by factors unrelated to individual teaching effectiveness, such as, perhaps, student crowding in the law building.

3. Reliability and Response Rates. Researchers have found that reliability of SETs can be undermined by low response rates and small class sizes. Random factors, such as a student having a bad day, will have more impact in small samples and affect reliability. In addition, the likelihood that non-responding students are a non-random and unrepresentative sample further

---

[10] Angela R. Linse, "Interpreting and Using Student Ratings Data: Guidance for Faculty Serving as Administrators and on Evaluation Committees," 54 Studies in Educational Evaluation 94, 96 (2017)

[11] Daniel E. Ho and Timothy H. Shapiro, Evaluating Course Evaluations: An Empirical Analysis of a Quasi-Experiment at the Stanford Law School, 58 J. Legal Educ. 388, 405 (2008)

[12] Bob Uttl, et al. "Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related," 54 Studies in Educational Evaluation 22, 23 (2016).

[13] Rice podcast.

[14] See Appendix 3.

suggests that low response rates produce unreliable numerical data. Therefore, advocates of numerical SET data have concluded that class means should not even be calculated in classes with fewer than 10 students and that a response rate of at least 66% is necessary for reliability.[15]

A debate in the literature persists over whether paper or electronic SET forms are preferable, and one of the issues is that response rates are lower with electronic forms in the absence of incentives or penalties to induce students to submit them.[16] UW Law's experience demonstrates this problem. The average response rate for paper SET forms at UW Law School from 1999 through 2007 was 82%. When the Law School switched to electronic SETs in the Spring 2008 semester, the response rate immediately dropped to 71%, and the average since then has been 67%, just over the minimum threshold for reliability. Electronic response rates have varied dramatically, from 56% to 77%, never even reaching the *lowest* response rate of the paper years (79%).[17] More worrisome is the fact that in six of the 22 semesters since Spring 2008, the Law School-wide SET response rate fell below the 66% reliability threshold (most recently in spring 2014), suggesting that Law School means for those years should be thrown out. In another seven semesters, the Law School response rate fell between 66% and 69%.

Even when greater than the minimum for reliability, these response rates are not high. Moreover, a leading study on response rates found that non-responding students on average would provide lower SET scores.[18] This suggests that individual course means with response rates higher than the law school average response rate may be biased downward and not fairly comparable to the law school mean.

Finally, the response rates for qualitative comments tend to be about 1/3 lower than the overall response rates. This means that qualitative comments are forthcoming from a high of around 50% of enrolled students. And many of these are dashed-off one liners rather than detailed or thoughtful comments.

4. Distribution to students. The distribution-and-response protocol is described in detail in Appendix 2, but readers will be familiar with its general outlines. Students are given access to an online course evaluation form two weeks before the end of the semester and have until the last day of classes to complete it. The two-week window means that students can complete the form at various times and circumstances that can affect inter-rater reliability. Students who complete a SET form two weeks before the end of the course may not be ideally positioned to assess whether the course wraps up in a coherent or comprehensive fashion. To be sure, most students probably complete the form during the recommended in-class block of time set aside by the instructor. Yet even this is subject to variation that affects comparison between instructors. For example, researchers have found that providing students with food treats at the time SET forms

[15] Rice podcast.

[16] Meredith J. D. Adams and Paul D. Umbach, "Nonresponse and Online Student Evaluations of Teaching: Understanding the Influence of Salience, Fatigue, and Academic Environments," 53 Res. Higher Ed., 576 (2012).

[17] See Appendix 4. This description of the data adjusts the data for the 2013-14 academic, which appears to be an anomaly resulting from shifting responses from the fall to spring semesters. The reported figures were 95% for fall based on an anomalously low numerator (total # of student responders) and 56% for spring, based on an anomalously high numerator.

[18] Maarten Goos and Anna Salomons, "Measuring teaching quality in higher education: assessing selection bias in course evaluations," 58 Res. Higher Ed. 341, 355 (2017).

are distributed has a significant positive impact on evaluation scores.[19] Some, though not all, members of the Law School faculty do this, and the Law School has not identified the practice as an issue of concern for the reliability of SETs. Similarly, a member of this Working Group reports that his former practice was to give an emotional "benediction" speech immediately before giving students the in-class time-block for completing evaluation forms; his impressionistic belief is that such a speech, which he has not undertaken in every class, had a noticeable positive effect on his SET scores.

5. Conclusion. At this point it is worth remembering that the UW Law SET does not even purport to measure teaching effectiveness. Instead, the form purports to measure ***student feelings*** about teaching effectiveness – a related, but different and oversimplified question that avoids hard questions about what constitutes effective teaching. The administrative and interpretive problems discussed above exacerbate the inherent reliability problems of our SET form *even as a measure of student feelings*.

III. University and Law School Policies Relating to SETs

A. University.

University policy appears to leave the contents, procedures, and uses of SETs to the discretion of the divisions and schools. The most centrally-driven SET policy is that issued by the Board of Regents to the UW *System*, and thus may not be applicable to the Madison campus.[20] While stating that "the Regents believe that student evaluations are an important and useful source of evidence that should be explicitly considered in reaching judgments" about teaching effectiveness, the Regents' policy leaves considerable discretion to campuses and units about SET contents, and how they are to be interpreted and weighted in forming such judgments. Further, the Regents expressly acknowledge that "no single instrument or methodology can be identified that is clearly more valid or useful than another," and that "student evaluation must not be a substitute for direct peer judgment of teaching effectiveness through a variety of means such as observation of teaching, assessment of syllabi, examinations and other course materials, and evaluation of contributions to development and strengthening of departmental curriculum."[21]

The rules of the Social Sciences Divisional Committee, which oversees most Law School tenure cases, list "systematic surveys of student opinion" as just one of eight types of "appropriate" evidence of "effective teaching abilities." The Divisional Committee makes no

[19] Michael Hessler, et al., Availability of cookies during an academic course session affects evaluation of teaching, 52 Medical Education 1064, 1069 (2018).

[20] ABA Standard 315 requires the dean and faculty of a law school to conduct an "ongoing evaluation of the law school's program of legal education." While SETs might arguably be assumed a traditional element of such evaluation, the ABA does not expressly say so, let alone make suggestions about the proper contents of student feedback surveys.

[21] Regent Policy Document 20-2 (formerly 74-13), "Student Evaluation of Instruction," https://www.wisconsin.edu/regents/policies/student-evaluation-of-instruction/#UseofStudentEvaluationsforRetention.Promotion.andTenureDecisions:

stipulations as to what those student surveys should contain.[22] Moreover, in fall 2017, the Divisional Committee issued a "major … revision to the role that student teaching evaluations play in establishing excellence in teaching." The memo observed that SETs "have a patina of quantitative rigor while in fact being extremely noisy measures," that are "influenced by many factors that don't necessarily reflect teaching excellence," including race and gender bias. Thus, the Divisional Committee concludes that "while [SET scores] might be informative about change within one instructor over time, we will not overly scrutinize absolute numerical values or comparisons across faculty in a department." The "reduction on weight of student feedback is balanced by additional consideration of other evidence of excellence in teaching," such as self-reflection statements on teaching and peer evaluation.[23]

Nevertheless, other segments of the University are lagging behind the Divisional Committee's assessment of SETS. The Nominating Procedures for Faculty Distinguished Teaching Awards requires

> A summary of student evaluations of teaching, including numerical data (where applicable) and a representative selection of student comments (if available) for each course taught. … [T]he summary should include: (a) a copy of the questionnaire(s) used showing the specific questions asked; (b) a description of how the student evaluations were solicited and administered; (c) the mean scores for each question asked in the evaluation of every class taught by the nominee; and (d) the department's mean scores for each question asked in the evaluation. [24]

The Law School will need to determine whether to use the current SET form for this purpose, or instead try to assume a leadership role within the University to reform traditional SET procedures.

B. Law School

1. Tenure and Promotion. When the Law School's most recent tenure dossier attempted to de-emphasize SET score means for the tenure candidate, the Divisional Committee gave feedback suggesting that doing so was inconsistent with the Law School's own policies. The Law School's tenure rules require demonstration "that the tenure candidate has become, and will continue to be at minimum a capable and highly motivated teacher."[25] The Tenure Rules provide further that

---

[22] Social Sciences Divisional Committee, Statement of Criteria and Evidence for Recommendations RegardingTenure https://secfac.wiscweb.wisc.edu/wp-content/uploads/sites/50/2018/09/Tenure-guidelines_SocSci_changes-highlighted.pdf

[23] Fall 2017 Memorandum from Maryellen MacDonald, Chair, Social Sciences Divisional Committee, attached as Appendix 5.

[24] https://secfac.wiscweb.wisc.edu/wp-content/uploads/sites/50/2018/09/2019-Nominating-Procedures-for-Faculty-Distinguished-Teaching-Awards.pdf

[25] University of Wisconsin Law School, Rules for Tenure Decisions, and Pre-Tenure Review of Untenured Faculty, Rule 2.1(a)(2), at 3-4.

> Teaching quality can be judged on a variety of measures, including student evaluations, peer evaluations, and evidence of effort, activity and innovation observed by members of the law school community or provided by the candidate. Numerical student evaluations of overall teaching quality at or above the faculty mean can be evidence of excellent teaching, though this evidence is not essential in all cases or necessarily dispositive.[26]

Finally, "Appendix A: Tenure standards vocabulary," states that "Inadequate teaching would ordinarily be evidenced by negative student or peer evaluations," whereas "[excellent] Teaching must be consistently effective and earn positive evaluations by students and peer teachers."[27]

2. Publication to Students. The practice of the Law School for the past several years has been to publish SET scores to students. SET scores are published on each faculty members' web page (except for untenured faculty), and aggregated on a law school web paged aimed at students, called "Summaries of Student Course Evaluations and Grade Distributions." It is the first link on the "Curriculum Guide" for students page: https://law.wisc.edu/academics/curriculum-guides/. That page displays this "warning":

> Student evaluations are an imprecise measure of faculty effectiveness. Research has shown that factors such as gender and race can affect students' perceptions of faculty. Likewise the size of the class and the difficulty of the subject-matter can affect students in the evaluation process. We are providing the numerical data from recent student evaluations at the request of the SBA, but they should be recognized as one of many sources of information about courses. Due to the difficulty of making generalizations from small groups, information about classes with less than 5 students, clinical courses, and independent readings courses is not available to students. Information about classes taught by tenure-track faculty who are not yet tenured is also not available to students.

The Committee has been unable to find any record of a faculty discussion or decision on this practice or the "warning" message. The Committee is concerned about the propriety of posting faculty SET scores as advice to students, in full knowledge that the information is "imprecise" and biased by race, gender and class size. Even with the above disclaimer, the posting practice puts the Law School's official imprimatur on bad information. There is no evidence that SET scores are more valid and reliable than informal faculty reputations passed by word-of-mouth among students.

IV. Academic Research on SETs

The desirability and utility of teaching evaluations is the single most studied question in the field of higher education. Since the first documented use of SETs in the 1920s, thousands of academic studies have been published on this topic.[28] Although this Committee has admittedly only scratched the surface of this research, we have been able to discern that the research falls

[26] Rule 2.1(b)(2), id., at 5.

[27] Id. at 25.

[28] Herbert W. Marsh, "Students' Evaluations of University Teaching: Research Findings, Methodological Issues, and Directions for Future Research," 11 Int'l J. of Educ. Res. 253, 260 (1987); Rice podcast.

http://cte.rice.edu/blogarchive/2015/02/01/studentratings

into three groups. (1) Zealous defenders of SETs continue to argue that properly-constructed and administered instruments can produce valid and reliable measures of teaching effectiveness, student learning, and student attitudes.[29] (2) Moderate defenders contend that SETs are properly used only as "student perception data," that are indirect or incomplete evidence of teaching effectiveness and student learning, but should not be taken as direct measures of either teaching effectiveness or student learning.[30] (3) Critics of SETs argue that they are severely and perhaps incurably unreliable, invalid, and race- and gender- biased as measures of teaching effectiveness and student learning.[31] There is thus an unsettled academic dispute over whether SETs are useable under any circumstances.

We do not need to settle this dispute in order to move forward, because there is no significant research that would defend the 5-7 likert-question SET form in use at UW and most other law schools. While critics contend that even the best SET forms suffer from race and gender bias, those who defend SETs rely on studies based on validated instruments designed by educational survey experts.[32] Even the most zealous advocates of SETs caution against use of "homemade" set forms like UW Law School's that are generated by faculty/student committees without reference to psychometric testing expertise.[33] Such SET forms as ours that "are developed without any clear theory of effective teaching" raise serious problems of validity, reliability, and bias.[34]

A. Reliability

Reliability of teaching evaluations refers to the consistency of an individual student's own rating from one SET form to the next, to consistency between student raters, and to the consistency of class averages from one course to the next. The research shows that individual student ratings have very low reliability. They are highly subject to extraneous factors and day-to-day variation.[35] Class averages have higher reliability, but with significant caveats.

First, the research supporting reliability of numerical SET scores is based on validated expert-created instruments. These consist of anywhere from 30 to 100 questions designed to create cross-checking and confirmation of answers through repetition. [36] (The recent faculty benefits survey administered by UW is an example of a survey instrument designed to increase

---

[29] Raoul A. Arreola, Developing a Comprehensive Faculty Evaluation System 98-124 (2007); Herbert W. Marsh, Distinguishing Between Good (Useful) and Bad Workloads on Students' Evaluations of Teaching, 38 American Educational Research Journal 183 (2001); William E. Cashin and Ronald G. Downey, Using Global Student Rating Items for Summative Evaluation 84 J. Ed. Psych. 563 (1992).

[30] Angela R. Linse, "Interpreting and Using Student Ratings Data: Guidance for Faculty Serving as Administrators and on Evaluation Committees," 54 Studies in Educational Evaluation 94, 95-96 (2017).

[31] See Uttl, et al., supra; see also sources cited in the next three subsections.

[32] Rice podcast.

[33] Areolla, supra, at 99-101.

[34] Pieter Spooren, et al., "On the Validity of Student Evaluation of Teaching: The State of the Art," 83 Review of Educational Research 598, 602-03 (2013); Rice Podcast, supra. For examples of SET defenders relying on expert-created instruments, see Marsh, "Students' Evaluations," supra, at 263-64; Cashin, supra, at 565.

[35] Rice Podcast.

[36] Areolla, supra, at 112-15; Rice podcast.

reliability through such repetition.) These SET instruments tend to use descriptive rather than evaluative scales and seek information about student experiences rather than broad opinions.

Second, even where such instruments are used, it is not clear that they generate reliable comparisons among professors or with a global mean in the Law School setting. Studies relied on by SET proponents tend to be based on large undergraduate courses comparing multiple sections of the same course, thus minimizing the effect of differences in subject matter, required versus elective courses, class size, and the like. UW Law School's database, with relatively small samples and large variation among course types and sizes, may simply be unsuited for comparative averages.

Third, as explained above, reliability is highly correlated to class size, which should not be surprising to the statisticians among us. Sufficiently large sample sizes can minimize the extraneous influences affecting individual responses. The research shows that classwide means are unreliable where there are fewer than ten responses, or in classes of any size with a response rate below 66%.

B. Validity

A necessary pre-condition to validating a test or survey instrument is to specify the object of measurement. SETs run into difficulties here, and the UW Law SET form is no exception. Without a clear faculty consensus on what elements make effective teaching, it is impossible to create a valid SET instrument, since validity by definition means successfully measuring an identified object.[37] There is no consensus about whether even the best SET forms can validly measure teaching effectiveness or student learning. A recent meta-analysis of past studies claims that there is no positive correlation whatever between SET scores and student learning, and indeed suggests that the two may be negatively correlated.[38]

Once the object of measurement has been determined, validating the SET instrument requires identifying one or more alternative measures of the object. For example, if the object is teaching effectiveness, measures such as self- and peer- teaching assessments or alumni surveys might be used. But given the lack of consensus that SETs can validly measure teaching effectiveness, some researchers suggest that the focus of student course surveys should place emphasis on objective facts about what goes on in the course, along with questions emphasizing the students' responsibility for their own learning, such as how much time the student spent studying outside of class, whether they discussed course material with fellow students, whether they read all of the assignments, and the like.[39]

The Rice University Center for Excellence in Teaching, which positions itself as a proponent of SETs, provides an instructive outline of best practices. Validated SET instruments must statistically control for subject matter, class size, and student characteristics (interest, work ethic, effort) *before* reporting results. Even carefully constructed, validated and administered instruments that produce valid (and even possibly the best) measurements of teaching effectiveness, are far from perfectly valid. Therefore, summative evaluations of teaching must

[37] Stark and Freishtat, supra, at 13; Spooren, at al., supra, at 602-03; Rice podcast, supra.

[38] Uttl, et al., supra.

[39] Rice podcast

include multiple measures, such as peer reviews and self-assessments. Evaluators must be trained to interpret the results of these complex instruments correctly. This includes knowledge and understanding of variables, lack of validity of decimal points, etc.[40]

UW Law has relied on "homemade" and un-validated SET forms since before 1973. The law of disparate impact only began to be developed in 1971 and the applicability of federal antidiscrimination law to state institutions began in 1972.[41] It is therefore doubtful that the legality of the course evaluation form was rigorously assessed at its inception, and there is no evidence of such an assessment in the faculty records. A test instrument that has a disparate impact on race or sex in employment is presumptively invalid, and generally requires evidence of validity in the form of a professionally-conducted validation study.[42] It is doubtful that the UW Law SET form could be validated, and this raises serious questions about the legality of relying heavily on the form in decisions regarding tenure, promotion, and performance evaluation.

C. Use of Global Questions

Question 5 on UW's SET form, asking about the instructor's "overall teaching quality," is what is known in the literature as a "global question." Inclusion of a global question seems to be *de rigeur*, yet there is a heated debate in education research about whether such questions should be used in SET forms to measure teaching effectiveness. Even SET proponents disagree about the reliability and validity of global evaluative questions.[43] Critics argue that reducing teaching effectiveness to one or two global questions is improper, as such questions are unreliable, invalid and possibly illegal.[44]

Because students are explicitly (as in the pre-2002 UW form) or implicitly permitted to consider Q5 as a stand-alone rather than a weighted average of the previous questions, the question invites unguided subjectivity and opens the door to race and gender bias.[45]

---

[40] Rice podcast.

[41] Griggs v. Duke Power Co., 401 U.S. 424 (1971); Equal Employment Opportunity Act of 1972, Pub. L. No. 92-261, § 2, 86 Stat. 103, amending 42 U.S.C. § 2000e-1 (1972).

[42] See, e.g., Joseph A. Seiner and Benjamin N. Gutman, *Does Ricci Herald a New Disparate Impact?*, 90 B.U.L. Rev. 2181, 2209-12 (2010).

[43] Compare Cashin and Downey, supra (defending global questions), with Marsh, "Students' Evaluations," and Marsh, "Distinguishing," supra (arguing that SETs must be "multidimensional" and not reliant on a single global rating),

[44] Ronald A. Berk, Should Global Items on Student Rating Scales Be Used for Summative Decisions?, 27 J. Faculty Devel. 57, 58 (2013).

[45] Amy L. Wax, *Discrimination in Accident*, 74 Ind. L.J. 1129, 1137 (1999); Melissa Hart, *Subjective Decisionmaking and Unconscious Discrimination*, 56 Ala. L. Rev. 741 (2005).

D. Race and Sex Bias

A powerful consensus in the literature has concluded that SETs are significantly affected by race and sex bias.[46] As summed up in a 2016 study, SETs are biased against female instructors to an extent large enough to cause more effective female instructors to get lower SET scores than less effective male instructors; SET scores measure gender bias better than they measure teaching effectiveness.[47] An internal study at UW Law School concluded that the sex of instructors had a statistically significant negative effect on SET scores, with women professors receiving an average of 0.19 lower scores on the Q5 mean than men, controlling for class size, 1L classes, and other factors.[48]

The problem of bias is exacerbated by the use of vague and subjective evaluative questions on SET forms. There is no reason to suppose that students are better able to determine the elements of effective teaching than our faculty, which has so far failed to answer that question in the context of the SET form. It is well-established in the literature of cognitive psychology that respondents faced with an inordinately difficult question will resort to heuristics, and that a commonplace set of heuristics involves racial and gender stereotypes and expectations that can lead to unconscious bias.[49] Students apply sex-role expectations to their instructors, and look for different mixes of personality traits: for example, students expect female instructors to be "personable and accessible," and give higher ratings to female professors coded as "androgynous."[50] Studies have also shown that SET scores are heavily influenced by superficial judgments based on non-verbal behaviors, cues, and appearances. One study showed that undergraduate students' "assessments" of instructors based on *a 30-second video clip without sound* were highly correlated with end-of-semester SET scores.[51] Several studies of SETs have shown that "global" assessment questions, such as Q5 ("overall teaching quality") on the UW SET form "are not strongly related to more specific measures and are more highly correlated

[46] Friederike Mengel, Jan Sauermann and Ulf Zölitz, "Gender Bias in Teaching Evaluations," IZA Institute of Labor Economics Discussion Paper Series (2017); Roxanna Harlow, "Race Doesn't Matter, But . . . : The Effect of Race on Professors' Experiences and Emotion Management in the Undergraduate College Classroom, 66 Soc. Psy. Q. 348 (2003); Heather Laube, et al., The Impact of Gender on the Evaluation of Teaching: What We Know and What We Can Do, 19 NWSA Journal 87 (2007); Lillian MacNell et al., *What's in a Name: Exposing Gender Bias in Student Ratings of Teaching*, 52 Journal of Collective Bargaining in the Academy 1 (2015).; Lisa L. Martin, "Gender, Teaching Evaluations, and Professional Success in Political Science," APSA Annual Meeting (August 29-September 1, 2013, Chicago, IL).

[47] Anne Boring, Kellie Ottoboni, and Philip B. Stark, "Student evaluations of teaching (mostly) do not measure teaching effectiveness," ScienceOpen Research (2016).

[48] Gwendolyn Leachman, Study of UW Law School Teaching Evaluation Data, February 6, 2018.

[49] Deborah J. Merritt, Bias, the Brain, and Student Evaluations of Teaching, 82 St. John's L. Rev. 235, 240 (2008).

[50] Laube, et al, supra, at 89-90; MacNell, et al., supra, 294-96.

[51] Nalini Ambady and Robert Rosenthal, Half a Minute: Predicting Teacher Evaluations From Thin Slices of Nonverbal Behavior and Physical Attractiveness, 64 Journal of Personality and Social Psychology 431 (1993). See also Dennis E. Clayson & Mary Jane Sheffet, Personality and the Student Evaluation of Teaching, 28 J. MARKETING EDUC. 149, 149, 157-58 (2006) (making similar findings comparing SETs given after five minutes exposure to the professor and again at the end of the semester).

with gender, status, and other contextual factors than are specific measures."[52] Yet even specific measures and qualitative comments are subject to race and gender bias.[53]

V. Peer Institution Practices

USC has made a university-wide decision, applicable to its law school, that SETs will no longer be used in tenure decisions. USC will continue to use a SET form emphasizing objective facts about teaching practices and students' self-assessment to help professors review their own course design and to shape the teaching reflection statements which are part of the review process.[54] USC appears to be at the forefront on this issue, though other universities are apparently engaged in serious re-evaluation of the SET processes.[55]

The UW Law School Ad Hoc Committee on Course Evaluations completed interviews with seven peer law schools: Berkeley, Marquette, Michigan, Minnesota, Northwestern, Ohio State, and UCLA.[56] One Working Group member interviewed the associate or assistant dean responsible for administering the SETs at each school, using a structured questionnaire developed by the Committee. (See Appendix 6).

The peer school interviews suggest that UW Law School is neither leading nor lagging in its approach toward SETs, but that *we place greater reliance on them and use them for more purposes than most of our peer institutions*. Most of the peer schools used SET forms similar to ours, with between 5-7 questions using likert scales of 5-6 points, plus an open-ended comments question. (UCLA uses 7 and 9-point likert scales on its form.) Berkeley's SET form employs 10 questions, and Michigan's form includes 12 likert-scaled questions that are required university-wide. All the peer schools include 1 or 2 "global" evaluative questions that are not tied to the more detailed questions. Like UW, none of the interviewed peer schools have a clearly articulated set of measurement objectives that would permit validation of the forms, other than "teaching effectiveness" or "what is on the form itself." In other words, all the interviewed peer schools appear to use unvalidated "homemade" forms.

There was more variability in the peer school's use and distribution of the SET data. Only one peer school (Ohio State) said that SET scores "feature prominently" in tenure decisions, a description that likewise could be fairly applied to UW Law. Minnesota reported that peer teaching reviews are weighted as or more highly than SETS. UCLA uses SET scores only if they fall in the lowest tenth percentile. Northwestern does not require tenure committees to look at SET means, although the associate dean provides them; student qualitative comments are not used in tenure files. Berkeley and Michigan place greater emphasis on qualitative comments than SET scores. While Marquette does use SET scores in tenure decisions, they break down the averages by race, sex, class size, and 1L courses, "to provide context."

---

[52] Heather Laube, et al., "The Impact of Gender on the Evaluation of Teaching: What We Know and What We Can Do," 19 NWSA Journal 87, 89 (2007)

[53] MacNell, et al., supra, at 8; Leachman memo, supra.

[54] "USC nixes student evaluations as part of tenure review," https://www.educationdive.com/news/usc-nixes-student-evaluations-as-part-of-tenure-review/524163/ ;

[55] See, e.g., https://provost.uoregon.edu/revising-uos-teaching-evaluations.

[56] An eighth, Iowa, was unresponsive despite several requests for an interview.

UW Law relies on SET scores informally in forming community reputations about teaching ability, and formally in nominating its faculty for University-wide teaching awards. Committee interviewers discussed community teaching reputations in four of the seven peer-school interviews. The associate deans at Ohio State, Michigan, and Berkeley suggested that faculty did not view SET scores as highly correlated to teaching effectiveness; the Northwestern associate dean said that while most faculty accepted that they were biased and unreliable, "they still like to use them, like placements in law reviews."

All of the peer schools calculate means of the likert-scaled questions, and the means all tend to cluster toward the high end of the scale.[57] But there the similarity among peer schools ends. None of the peer schools appear to push out numerical SET scores with comparative faculty-wide means, as UW Law does. Ohio State and Marquette withhold SET data from students entirely. Michigan declines to "advertize" the availability of faculty SET scores to students, though it makes them available "on request" in the registrar's office. Berkeley posts individual faculty SET scores, and Northwestern posts both SET scores and qualitative comments, to assist students in selecting courses, but neither provide faculty-wide means. Minnesota is required by state law to post the percentage of students who say on the SET form that they "would recommend the course," but this does not appear to include comparison to a faculty mean. Northwestern and Ohio State do not report faculty-wide means to faculty members. Berkeley breaks down faculty-wide means by class size, Minnesota separates 1L and upper-level course means, and Marquette, as noted, sorts by class size, 1L versus upper level courses, race and gender. Associate deans at Minnesota, Michigan, and UCLA use SET scores to look for outliers and those, along with Northwestern, use SET scores to identify sub-par adjunct professors. The associate dean of Northwestern reviews qualitative comments to identify issues of race or sex bias in the classroom.

Several of the peer schools indicated concerns about bias or unreliability of their SET forms. Berkeley is "not rethinking the enterprise," even though the faculty appears to have "given up on" reliability and validity. But Marquette, Northwestern, Ohio State, and UCLA are all at some stage of reassessing their SETs. Marquette, Northwestern and Ohio State have campus-wide efforts in place to revise the SET process due to reliability-validity-bias concerns. The Northwestern associate dean believes their SET form shows race and gender bias based on her own analysis. At Marquette Law School, there are informal conversations about changing the SET process, and UCLA is looking at changing its form to better accommodate experiential learning courses.

VI. Recommendations

The Committee makes the following four recommendations:

1. **Reporting of SET scores.** The Law School should indefinitely suspend reporting numerical SET scores to faculty and students, effective Spring semester 2019 until the implementation of a valid and reliable long-term SET protocol (see recommendation 4). SET score reports currently made available to students should be removed from the Law School

[57] See, e.g., Ho & Shapiro, supra, at 394-95.

web site. The Law School should continue to collect the numerical data for purposes set out in recommendation number 2.

2. **Use of SET data.** On an interim basis, pending the implementation of a valid and reliable SET procedure (see recommendation 4), the use of numerical SET scores for tenure, promotion, and performance evaluation decisions should be suspended, effective immediately. Qualitative comments should be distributed to instructors for their use in self-assessment of teaching. Qualitative and numerical data from the current SET form should not be used as conclusive evidence of teaching quality, but may be considered as an indicator of need for further supervisorial inquiry into classroom problems. Qualitative and numerical data from the current SET form may continue to be used for determination of adjunct professor renewal. All uses of any data from current SET forms should be with due regard for their reliability, validity, and bias problems.

3. **Revision of applicable Tenure Rules.** The Law School's Tenure Rules should be promptly amended to conform to recommendation 2.

4. **Standing Committee to revise the SET process.** A standing faculty committee on SETs should be convened. As a first order of business, it should create a new interim evaluation form within the next year. The form should emphasize student self-assessment and objective facts about teaching practices rather than subjective teacher evaluation. The standing committee on teaching evaluation should thereafter study and propose a teaching evaluation system, and a reliable and valid student feedback protocol for long-term use, based on consultation with experts in the field. The standing committee should expeditiously determine whether and how SET data should be used in nominating law school faculty for university teaching awards.

A. Recommendations 1 and 2: Restricting Use of SET Data

The Committee believes that the current SET form is so fraught with reliability, validity, and bias problems that its use should be severely restricted. At least two members of the Committee believe the form should be discontinued full stop. Others on the Committee prefer a compromise approach to avoid leaving the Law School without any student feedback mechanism during an interim period in which a valid and reliable SET protocol is studied and implemented. The recommendations thus reservedly embrace certain second-best or least-worst alternatives

1. Suspend Reporting of SET Scores and Averages. The use of likert numbers exacerbates the reliability, validity, and bias problems inherent in the UW Law SET form. The lure of numbers, particular the reduction of teaching effectiveness to a single mean for Question 5, is extremely powerful, suggesting (falsely) the ability to compare teaching ability among facuty members in a simple and cheap way. The use of numerical SET averages implies the Law School's belief that the Law School form is reliable and valid. The use of numbers implies validity because the numbers are treated as a measurement of an identified object: teaching effectiveness. Yet the Law School has no identified, articulated consensus on what effective teaching consists of. The use of numbers implies reliability by reducing every instructor's evaluation for every course to a single number that is treated as fully commensurable with the number generated by every other course and with a schoolwide mean. By publishing numerical averages and law-school-wide means, the misuse of these numbers for unduly simple and easy

comparative purposes is almost irresistible, and the numbers are in fact used as though they were valid apples-to-apples comparisons.

If, as the Committee believes, the form is in fact neither sufficiently reliable nor valid, the reliance on numerical averages for comparative purposes merely serves to "double down" on the SET form's deficiencies. Because of the form's dubious reliability and validity, and its susceptibility to illegal bias, it should not be used in performance evaluation or tenure decisions in most circumstances.

The reporting of SET numbers is unduly demoralizing for faculty members whose teaching in fact may be good or very good but whose Q5 averages fall below the law school mean. The current heavy reliance on the Q5 number allows unstructured and probably biased student opinions to carry undue weight without supporting reasons.

By generating misinformation about teaching effectiveness, SET numbers are misleading for students' course selection decisions. Reporting them to students is thus inappropriate, and may cause "echo chamber" effects on instructors' future SET scores.

2. Restrict Use of SET Data. The Committee recognizes that SET scores may need to be relied on in two limited circumstances: to make decisions regarding rehiring of adjunct instructors and to identify possibly extreme underperforming teachers. Whereas faculty hires are vetted much more intensively at the front end and are supervised by faculty processes, adjuncts are hired and supervised with far fewer controls, by one or two administrators without faculty oversight. Some kind of tool is needed, at least in the interim, to identify underperforming adjunct instructors. Dean Kelly reports that at present he relies on the SET form for this purpose.

Course means falling below the midpoint of the likert scale (3.0) can alert the Dean's office to the need to make further inquiry into an individual instructor's significant teaching problems. For that reason, the Committee recommends that SET scores continue be calculated even though not distributed or used for other purposes.

Qualitative comments are also subject to reliability, validity and bias problems, and their use, too, should be curtailed. However, the Committee was not persuaded that these problems required a complete suspension of any student-to-teacher feedback mechanism. Therefore, the Committee recommends that qualitative comments continue to be distributed to instructors for their use in self-assessment of teaching. It is worth noting that problems with qualitative comments are not as extreme as with numerical ratings. The numbers create an aura of precision and inter-rater reliability that we do not associate with qualitative comments. Moreover, the numbers are opaque, whereas a qualitative comment at least attempts an explanation. Biases and other teaching-irrelevant factors will be at least somewhat easier to detect from the language used. It is easier to sort out rants and negative outliers from a judgment about overall teaching quality than the numbers 3, 2 or 1 factored into a classwide mean.

B. Recommendation 3: Tenure Rules and Performance Evaluations

The Committee recommends the following Tenure Rule changes, indicated in strikeouts for proposed deletions and bold for proposed additions:

> 2.1 (b)(2). Teaching quality can be judged on a variety of measures, including ~~student evaluations,~~ peer evaluations, **self-assessment,** and evidence of effort, activity and

innovation observed by members of the law school community or provided by the candidate. Numerical student evaluations **based on the Student Evaluation of Teaching form in use at the law school as of 2018 shall not be used as direct evidence of teaching quality.** ~~of overall teaching quality at or above the faculty mean can be evidence of excellent teaching~~, **Qualitative student comments on that form shall be consulted,** though this evidence **of teaching quality shall not be given more weight than any other type of evidence, and due regard shall be given to the possible influence of race and gender bias.** ~~is not essential in all cases or necessarily dispositive.~~

Appendix A: Tenure standards vocabulary.

… Inadequate teaching would ordinarily be evidenced by negative peer evaluations, ~~and~~ coupled with a failure to respond to intervention and counseling by those charged with planning the curriculum~~, coupled with a failure to respond to that counseling~~. **Negative student evaluations must be clearly established through qualitative comments and corroborated by other evidence to establish inadequate teaching, and evaluators must take account of possible improper biases.** … [Excellent] Teaching must be consistently effective ~~and earn positive evaluations by students and peer teachers~~ **as shown by the types of evidence described in Rule 2.1(b)(2)**.

C. Recommendation 4: Amending the SET form, Interim and Long-Term

The Committee believes that a long-term SET protocol that is reliable and valid will require significant study and consultation with experts in the design of student feedback surveys. This may take in excess of one year of committee work, and may require ongoing monitoring. For that reason, the Ad Hoc Committee recommends establishing a standing committee on SETs.

Because the current SET form is so deeply flawed, the Committee further recommends that a newly-convened standing committee expeditiously attempt to substitute an improved SET form for interim use. Such a form should pose questions intended to obtain specific teaching-relevant information geared toward what students are well-positioned to report: facts about teaching practices they observe, and self-assessment relevant to their learning. Evaluative questions could be more specific, and less vague or global. While such an interim form may not be an adequate substitute for a professionally-designed and -validated form in the long run, it would represent an improvement over the current SET form and initiate a useful faculty-wide dialogue on what effective teaching is.

In the longer term, the standing committee should more exhaustively evaluate the options for a future SET procedure based on best practices. What are the goals of the form and objects of measurement? Should experts be consulted in constructing a new SET form and procedure? What do faculty, administrators and students want from a student feedback procedure? These questions require extensive information-gathering beyond the scope and capacity of an ad hoc committee.