The Place of Student Evaluations in Assessing Faculty Teaching

A report by the Albany Chapter of United University Professions
Prepared by Aaron Major
Assistant Vice President for Academics

October, 2016

The University at Albany aspires, quite rightly, to deliver the highest quality instruction to its students. Successful teaching depends not only on our faculty's diligent commitment to classroom practice, but also on the systems in place to formally evaluate the quality and effectiveness of faculty performance. The UUP Contract with New York State is very clear on the subject of evaluation: academic faculty can only be evaluated by other academic faculty. This report starts from the premise, codified in our contract, that any University procedure to evaluate the teaching effectiveness of academic faculty must be based on a system of peer-review.

Over the past several decades UAlbany's system of assessing academic faculty teaching has placed increasing—we argue *undue*—emphasis on student "evaluations" of our teaching via the Student Instructional Rating Form, or SIRF. Academic faculty, however, are skeptical of SIRF assessments. This growing skepticism is driven by, on the one hand, widespread attention given to several recent, high quality research studies that cast serious doubt on the validity and reliability of such student evaluations of teaching and, on the other hand, our University's shift to an on-line system of administering the SIRF, which has resulted in low student response rates that potentially give disproportionate weight to the views of a few disgruntled students. An extensive analysis of University at Albany SIRF data conducted in 2010 by the Course Assessment Advisory Committee shows that academic faculty are right to be skeptical of the instrument's validity as measure of teaching effectiveness. SIRF scores are systematically biased by low response rates, which have become much more common with the shift to online SIRFs, are biased against female instructors, and, because there is a strong, positive correlation between SIRF scores and students' expected grades, create pressures for grade inflation.

The Albany Chapter of United University Professions works to ensure that all of our members are treated fairly as employees of the University. Too often, we have found, career-defining decisions about renewal, promotion and tenure are made on the basis of the deeply-flawed SIRF. This is particularly the case for our contingent academic members for whom the SIRF is very often the only formal evaluation that they receive. In many Colleges, Schools, and

Departments, SIRFs function, we contend, as the *de facto* system of evaluation of faculty teaching. For many of our contingent faculty, for whom teaching constitutes their entire professional obligation, results from the SIRF constitute a *de facto* evaluation of their entire academic job performance.

Our University's current reliance on the SIRF to evaluate faculty teaching is, to put it plainly, bad for faculty, bad for students, and bad for the University as a whole. To the degree that the University relies on SIRFs as "student *evaluations*"—and we believe that department chairs, deans, university committees frequent do just this—it violates the contractual obligation that only academic faculty can evaluate other academics. The increasing weight that is placed on SIRF scores in reviews of academic faculty for renewal (especially for contingent faculty), tenure, and promotion is deeply troublesome for three main reasons. First, the over-reliance on student evaluations of faculty teaching is contrary both to contractual obligations on the evaluation of academic faculty and the existing University Senate legislation governing the evaluation of faculty teaching, which de-emphasizes the importance of the SIRF. Second, the reliance on student evaluations of faculty raises questions of basic procedural fairness as student SIRF scores are biased against female faculty and faculty who teach large courses. Third, over-reliance on the SIRF is a disservice to students who would benefit from a substantive, multi-faceted evaluation of faculty teaching by their expert peers.

Background: the shift to the on-line SIRF and the CAAC report.

To address growing concerns regarding the use of student evaluations in the review process for renewal, promotion, tenure, and reappointment, in 2010 the Provost charged the Course Assessment Advisory Committee (CAAC) to evaluate the University's course evaluation procedures and tools. The context for this initiative was the effort to shift the evaluation process from in-class, paper SIRFs (which were costly both in the materials required and the use of the University's test-scanning facilities) to an online form that students would complete during a designated window at the end of the semester. As part of its report, the CAAC conducted a statistical analysis of student evaluations between 2005 and 2010 and published these results as part of its 2012 Report of the Course Assessment Advisory Committee.[1]

To date, this report remains the most thorough and systematic analysis of the reliability and validity of the SIRF tool as it has been administered at the University at Albany and, equally important, stands as the most recent effort by the University to give thoughtful, careful consideration of the place of student assessment of faculty teaching within existing University and Senate policies regarding the evaluation of teaching. Given growing concern about the fairness and usefulness of student evaluations of faculty teaching, it is time to revisit the

[1]The final report of the committee can be found here: http://www.albany.edu/ir/CAAC%20FINAL%20Report.pdf

CAAC report on student evaluations and, in so doing, renew the conversation about fair, effective evaluation of faculty teaching.

The Committee's overall finding was that student evaluations are an imperfect, but nevertheless useful and valid instrument upon which to base formative and summative evaluations of faculty teaching. A careful review of the Committee's report suggests that even this hedged conclusion does not correspond with the data it analyzed. The Committee's own analysis of five years of SIRF scores shows that student evaluations are biased against female faculty, biased by response rate, punish faculty who teach large classes, and reward faculty for giving out higher grades. That these factors significantly influence student evaluations of faculty shows that the instrument is not a valid one, especially for making important, career-shaping decisions about renewal, tenure, and promotion.

Data from the CAAC report

The analysis in the CAAC report draws on five years of University at Albany SIRF score data from fall 2005 to spring 2010, creating a dataset of the full population of 319,320 individual student evaluations over that period. The data is pooled across semesters, and separate regression analyses were run for those evaluations given during class time on paper, and those evaluations that were administered online. While the SIRF asks students to rate their instructors and courses across several measures, the CAAC's regression analysis focused on the overall instructor rating. Several measures of the characteristics of the students (gender, year in school, average GPA, grade expected in class) and the courses (size, level, time of day of class meeting) were regressed on this dependent variable.

The table below reproduces the results of the CAAC's regression analysis[2] reporting the standardized and unstandardized regression coefficients and their significance. The variables are sorted according to the size of the standardized (Beta) coefficients, which provides a standard measure of the size of each independent variable's effect on students' overall instructor ratings. As the table shows, for evaluations given out on paper and in class, the five variables with the strongest effect on student evaluations of their instructors are 1) expected course grade, 2) class average GPA, 3) whether the course is a graduate level course, 4) the response rate of the evaluations for that class, and 5) the instructor's gender. For those evaluations administered online, expected grade, class average GPA, response rate and gender are still some of the more powerful predictors of how students evaluate academic faculty.

[2] These results can be found in the Appendix to the CAAC report, pages 22-26.

Results from the CAAC's analysis of SIRF data, 2005-2010: factors affecting
OVERALL INSTRUCTOR RATING (results sorted by size of Beta)

EVALUATIONS ADMINISTERED IN-CLASS	B	Beta
Expected Grade	**0.322**	**0.221**
Class average GPA	**0.244**	**0.113**
Level 500+	**0.304**	**0.111**
Response rate	**0.531**	**0.09**
Female instructor	**0.11**	**0.051**
Start 5:45 or later	**0.121**	**0.042**
Size 41-99	**0.092**	**0.038**
Start time: 9:20 - 10:00 a.m.	**0.157**	**0.034**
Size 150+	**0.093**	**0.032**
FT lecturer	**0.122**	**0.031**
Junior	**0.072**	**0.03**
Size 100-149	**0.079**	**0.026**
Non-Matriculated	**0.198**	**0.023**
Start 4:15 - 5:44 p.m.	**0.076**	**0.023**
Size 20-40	**0.049**	**0.022**
Instructor Sex code missing	**0.721**	**0.02**
Class meets minor	**0.058**	**0.018**
Start before 9:20 a.m.	**0.063**	**0.017**
Sophomore	**0.044**	**0.016**
Other instructor	**0.096**	**0.015**
Freshman	**0.041**	**0.015**
Level 300/400	**0.032**	**0.015**
Junior faculty	**0.036**	**0.014**
Female student	**0.026**	**0.012**
PT adjunct	**0.025**	**0.011**
Student sex code missing	**0.069**	**0.01**
Class meets major	**0.019**	**0.009**
Senior	0.011	0.005
TA instructor	0.01	0.002

EVALUATIONS ADMINISTERED ON-LINE	B	Beta
Expected Grade	**0.444**	**0.285**
Class average GPA	**0.23**	**0.095**
Size 150+	**0.272**	**0.092**
Level 500+	**0.211**	**0.074**
Size 41-99	**0.2**	**0.07**
Response rate	**0.245**	**0.042**
Junior faculty	**0.117**	**0.04**
Female instructor	**0.093**	**0.04**
Size 100-149	**0.084**	**0.026**
Start before 9:20 a.m.	**0.095**	**0.024**
FT lecturer	**0.098**	**0.022**
Size 20-40	**0.055**	**0.022**
Start time: 9:20 - 10:00 a.m.	**0.108**	**0.021**
Class meets minor	**0.07**	**0.02**
PT adjunct	**0.047**	**0.02**
Female student	**0.033**	**0.015**
Senior	0.025	0.009
Junior	0.023	0.009
Instructor Sex code missing	0.294	0.008
Level 300/400	0.017	0.007
Class meets major	0.016	0.007
Other instructor	0.016	0.003
Start 4:15 - 5:44 p.m.	0.009	0.003
Start 5:45 or later	0.009	0.003
Freshman	0.005	0.002
Sophomore	0.005	0.002
TA instructor	0.003	0.001
Student sex code missing	0.012	0
Non-Matriculated	0	0

*Coefficients in **bold** are significant at .05 or less.

What do these results suggest about the validity of the SIRF instrument as a tool for effectively evaluating faculty? In its report, the CAAC does draw attention to the finding that students who expect to earn a higher grade in a class evaluate faculty more favorably. Discussing this finding, the Committee notes: "the relationship between students' expected (or actual) grade and their ratings of instructors are potentially of interest in terms of the validity of ratings" (p. 12). While this finding by itself raises questions about the validity of the SIRF, more troubling is the report's silence on other factors.

In particular, the CAAC's data shows a strong effect from the response rate to the SIRF on evaluations; the lower the response rate, the lower an instructors' rating. Given that one of the Committee's charges was to specifically evaluate the validity of online evaluations, the Report's complete silence on the effect of response rate on evaluation scores is troubling. The Committee does suggest that low response rates (below 30%) should be 'viewed with caution" (p. 14). This would be an appropriate conclusion if the effect of low response rates were to increase random variability in evaluations. Yet the regression results show that low response rates are systematically biasing evaluations downward. This points to a negative response bias in evaluations—students who more readily complete evaluations are more likely to be those with negative reactions to the instructor—which also points to the invalidity of the SIRF evaluations.

A careful reading of the study therefore shows that student evaluations are not just an imperfect measure of instructor performance: they are an *invalid* measure of instructor performance. Student evaluations may measure a student's impressions of a course or an instructor, but they do not measure whether the instructor effectively conveyed the course material. This suggests a much stronger conclusion than the CAAC's Report offers, namely that students are not appropriate assessors of faculty teaching for summative purposes.

Student evaluations within the context of existing Senate policy regarding the evaluation of faculty teaching.

While the recent move to online administration of the SIRF has spurred renewed interest in the place of student evaluations in formative and summative judgments of faculty teaching effectiveness, this is not the first time that these issues have been taken up by the University community. In addition to performing a statistical analysis of SIRF results, the CAAC was also charged with providing a "summary of historical use and policy" of student evaluations in the teaching assessment process. The Committee's report outlines this legislative history on pages 4-5, excerpted below:

> The most important legislation governing assessment of teaching is Senate Bill 8384-07, implemented by administrative memorandum in April of 1984 and revised in 1991. By policy, student feedback is

> regularly solicited for courses, using the Student Instructional Rating Form (SIRF) coordinated by the Office of Institutional Research, or some other instrument endorsed by the instructor's department or program. The guidelines promulgated in 1984 specify that "all students shall be given an opportunity to make an evaluation in every class each term" and mandate that the collection of student opinions should be formulated and administered systemically at the department level (2005-06, p. 2).
>
> The 1984 Senate legislation went beyond student course evaluations. It also called for peer evaluation of teaching and noted that the methods for proper peer review had been less fully considered. To support peer evaluation of teaching, the legislation described several accepted techniques as examples for the faculty of each unit to use in developing a system that is tailored to the particular needs of their curriculum. It also mandated the use of peer evaluation in decisions concerning continuing appointment, but noted that departments were to be given "broad latitude" in developing systems for collecting and interpreting peer evaluations of teaching (2005-06, p. 1).

The CAAC committee interpreted the existing Senate legislation as follows:

> 1) that Senate Bill 8384-07 mandates that student evaluations be part of the assessment academic faculty's teaching.
> 2) that Senate Bill 8384-07 calls for peer evaluation to be part of the assessment process, but does not give peer evaluation any particular importance over other forms of assessment.

While the CAAC report correctly identifies the key pieces of legislation relating to student evaluations, it misconstrues the letter and spirit of that legislation. A fair reading of Senate Bill 8384-07 makes clear that the Faculty Senate intended to *reduce* the role of student evaluations in the assessment of faculty teaching by *making peer review the primary mechanism of faculty teaching evaluation and to subsume student evaluations of teaching under the peer review process*.

Senate Bill 8384-07[3]

[3] Page references to the Bill do not follow the Bill's own pagination (which does not list page numbers for the cover page and statement of background information); they reference the page number relative to the length of the entire document.

Senate Bill 8384-87 emerged out of a policy statement submitted by the Educational Policy Council (EPC)[4] to the Faculty Senate on December 5, 1983. A memo from Fred Volkwein to then President O'Leary dated December 6, 1983 suggests that the EPC policy statement was adopted in full. To the best of our knowledge, it remains the most recent Senate bill covering the evaluation of faculty teaching and is thus current Senate policy regarding evaluations of faculty teaching.

Contrary to what the CAAC report suggests, this bill does not establish policy with respect to student evaluations. As the bill describes in its own background summary, those policies had been put in place in previous years (1980 and 1981). Rather, Senate Bill 8384-07 emerged out of an effort to create a *comprehensive* policy for the evaluation of faculty teaching which, as stated in the bill "stresses the centrality of peer review in the evaluation of teaching (p. 2)." While it is difficult to know the intentions of these faculty senators in hindsight, the language of the bill strongly suggests that this effort was motivated by a desire to curb the influence of student evaluations of faculty by subsuming that process to an over-arching peer review process. The SIRF system had been put in place two years prior to this and the bill suggests that there was concern about how student assessment would factor into overall evaluations of faculty teaching. As stated in the bill in its opening page: "There are a number of guidelines on this campus regarding the collection and use of student opinion in the evaluation of teaching, but there is at present no comprehensive statement concerning the role of peer review (p. 2)."

Senate Bill 8384-07 contains language showing that the EPC was skeptical of the usefulness of "student opinion" and mindful of the limited interpretation that should be given to this form of review. As the Bill states, student evaluations were seen as providing a measure of student *perceptions* of teaching effectiveness, not a measure of actual teaching effectiveness. Moreover, the EPC drew attention to "studies showing a statistically significant impact of subject matter, class size, and course level upon student ratings of instructors" (p. 6).

The EPC's emphasis on peer review was grounded in a belief that "within the faculty resides the special competence needed to design the various programs of the curriculum, to make staffing decisions for courses, and to establish the standards by which student achievement is certified. *Primary use of that same competence must be made in evaluating teaching* (p. 3, emphasis added)." This principle, we believe, fully accords with the procedures for academic evaluation stipulated in the UUP contract.

To implement this general principle, the EPC charged academic departments with establishing "a credible and defensible method of evaluation of teaching (p. 4)." While the EPC did not make any recommendations for mandatory items to be included in each department's procedure, it did offer

[4] The Educational Policy Council is now the University Planning and Policy Committee (UPPC).

several suggestions. Items that could be submitted in support of such an evaluation—including syllabi, assignments, reading lists, grade distributions, and student questionnaires—were all to be submitted “in support of peer review (p. 5).

In other words, not only did the EPC not give student evaluations any more pride of place than any other item that could be entered into this portfolio, it also explicitly includes student questionnaires as one of many items that *could* be included, the decision of which items to include “decided upon in the context of each department’s procedures (p. 5).” Indeed, the Senate Bill also allows the departments “to transmit a *summary* of the student response data (p. 6, emphasis added),” rather than the complete results of student evaluations.

In setting up the context and background for its Report, the CAAC is right to hold up Bill 8384-07 as the critical policy statement on the evaluation of teaching. However it misconstrues the letter and spirit of Bill 8384-07. The CAAC’s Report interprets Bill 8384-07 to suggest that peer review serve alongside student evaluations as part of a ‘mixed method’ approach to faculty evaluation of teaching. A closer review of,the language of Senate Bill 8384-07 makes clear that student evaluations were, by policy, supposed to be subsumed to a process of peer review and given no special importance relative to other tools for assessing faculty teaching.

Moving forward: restoring the primacy of peer review.

Senate legislation establishing a framework for evaluation of faculty teaching asserts the primacy of peer evaluation and subsumes student review under a peer- and department-driven comprehensive evaluation procedure. Since the time when that Senate legislation was passed, evaluations of faculty teaching for formative and, most critically, summative purposes have become overly reliant on student evaluations. It is not uncommon for deans and chairs to make decisions about teaching effectiveness—including decisions of non-renewal—on the sole basis of SIRF scores. Such over-reliance is a) contrary to the letter and spirit of existing policy, b) as shown through an analysis of five years of UAlbany SIRF data, uses a highly imperfect instrument to guide decisions about teaching effectiveness, and c) does not comply with contractual obligations for academic review. There is no doubt that the academic faculty bears some of the blame for this current state of affairs. Comprehensive peer review processes are time consuming and while departments may be willing to devote the time and resources needed to conduct a more thorough evaluation of teaching for their tenure track colleagues, the growing ranks of part-time and contingent faculty across the University tend not to be evaluated in any manner except SIRF scores. This problem did not emerge overnight, and it will take thoughtful, careful discussion, planning and policy-making to correct it. In an effort to move that process forward, we suggest the following measures and principles.

1. The university needs to return to a process of evaluation of faculty teaching based on the principles laid out in Senate Bill 8384-07 and echoed in the UUP contract. Specifically, such a process needs to be peer-driven, shaped by the expertise that each department has over its subject matter, and open to a variety of tools and metrics. Those based on the evaluation of one's expert peers should take priority.

2. SIRF scores must no longer be used as the sole or primary measure of teaching effectiveness. While a more comprehensive evaluation procedure is developed, faculty should be able to freely choose between different student evaluations tools, such as narrative evaluations or those provided through the Institute for Teaching, Learning and Academic Leadership (ITLAL). In addition, faculty should never be evaluated on the raw results of student reviews, but rather should always be given the opportunity to reflect on those results through a narrative summary and should have those reviews analyzed and assessed by other academic faculty. In addition, faculty committees and administrators who play key roles in the review, tenure and promotion process should establish clear procedures and guidelines for how SIRF and non-SIRF student reviews of faculty teaching will be contextualized within a broader peer evaluation of teaching effectiveness.

3. We believe the University should end the practice of publicizing SIRF scores. Even with return rates above 60%, the scores are an invalid measure of teaching effectiveness, and retain biases against large classes of faculty members. Publicizing results legitimizes a deeply flawed practice.

4. We encourage the University Senate to review its existing policies setting the principles and procedures for evaluations of faculty teaching and, in so doing, further encourage the Senate to reaffirm the letter and spirit of Senate Bill 8384-07, particularly those elements that (a) assert the primacy of a flexible, peer-driven evaluation procedure and, (b) assert that each academic department has the "special knowledge" needed to develop their own evaluation procedures (but see points 5 and 6, below). In working to establish new principles and policies, the Senate should work closely with the UUP Albany chapter leadership, or its designees, to ensure that the system of evaluation of faculty evaluation comply with the letter and spirit of our collective bargaining agreement with the State.

5. We are sensitive to concerns that peer review can also be a highly imperfect instrument. Without clear guidance from existing best practices, peer review is likely to suffer from similar concerns about validity and reliability as student evaluations and thus subject faculty to similar kinds of

bias and discrimination. Peer review can also be problematic in cases where departmental relationships are contentious. Given that Senate Bill 8384-07 charges each academic department with developing its own evaluation procedures, departments would benefit immensely from guidance from existing best practices.

6. The process of evaluation of faculty teaching should be flexible to accommodate the professional standards and practises of various departments. However, all departments should establish procedures that are based on frequent consultation with the faculty member being evaluated, including:

    a. the ability of the evaluated faculty member to present and interpret their own portfolio of evidence of teaching effectiveness;
    b. the ability of the evaluated faculty member to review the presentation of their teaching record before it is sent for consideration by the larger faculty, and
    c. the ability to respond to outside evaluations of their teaching at all stages in the process.

    These procedures should be made clear as part of a readily-accessible document outlining the larger departmental procedures and policies for renewal, tenure and promotion. Finally, each department will need to develop clear processes for peer review of teaching that are applied to both the full-time and part-time faculty.

7. Currently, ITLAL provides a general set of guidelines and best practices for departments to consider as they develop their own peer review processes.[5] These resources are helpful, but could be bolstered by more concrete examples of peer review processes at peer institutions. In addition, ITLAL should be in a position to serve as an outside evaluator of faculty teaching.

We recognize that developing and implementing this kind of rigorous, fair, multi-dimensional and procedurally transparent evaluation process will require a great deal of time and energy from the University faculty. However, the importance of this issue demands such an investment. For the faculty, it is a matter of basic fairness as we move through the processes of renewal, tenure and promotion. For our students, it is a matter of ensuring that our curriculum is up to date and intellectually rigorous, preparing them not only for their future careers, but also to be thoughtful, engaged members of their communities. For the University, it is a matter of fulfilling the mission of SUNY to provide an education of the highest quality.

---

[5] ITLAL's Peer Observation Resource Book can be found at http://www.albany.edu/teachingandlearning/tlr/peer_obs/