



**United University Professions
Albany Chapter**

Labor-Management Notes

April 25, 2019

Management: J. Bochinno, W. Hedberg, R. Stark, H. Rodriguez, B. Szelest and B. Selchik

Labor: P. Stasi, A. Woulfe, L. Pyles, A. Major, T. Hoey., M. Seidel

1. Next meeting will be in May and no meetings until August.
2. Fees for undergraduate and graduate students: We are requesting information about fees for undergraduate and graduate students enrolling in courses at the University. Specifically, we would like to know:
 - a. Total University revenue attributable to such fees over the last five academic years.
 - b. Average annual fees charged per graduate student over the last five academic years.
 - c. Average annual fees charged per undergraduate student over the last five academic years.

-We are concerned about fees for several reasons including: the potential they have to erode our negotiated 'free tuition' benefit for our members; the undue burden that they put on graduate students, further eroding our University's competitiveness in attracting students; the undue burden that they place on undergraduates, undermining the University's commitment to supporting a racially and socioeconomically diverse student body. We would also like to take this opportunity to transmit a resolution, recently passed by our Executive Committee, supporting GSEU's efforts to reduce fees for graduate students [attachment 1].

-Labor is trying to get good data on what the fees actually are, particularly in terms of changes over time. Management noted that SUNY-wide has the fees available. They are also on the financial aid webpage. Finance put together broad-based fees from 2013 to present; it is currently \$2886 for undergrad and \$1908 for graduate. Student financial services is doing a breakdown. Can get us a summary once it is complete. Any changes

with fee structure go through student associations and SUNY and then is reviewed by the President. Todd Foreman is meeting with the graduate association to talk about fees, stipends, etc. Also, the issue of the timing of the fees versus stipends was forcing them to pay fees before they had their stipends.

3. Years of service awards for part-timers over the 2018-2019 academic year. Article 20.14(b) of the Agreement provides for a service award of \$500 for part-time faculty after 8 years of consecutive service to the University. We are requesting a list of eligible faculty. We are also asking what, if any, measures are taken to notify eligible faculty that they are receiving this award.

-Is there a protocol to do this? UUPs member data lags a bit so it's a bit hard to sort out. HR is going through info and trying to get data and will provide report. Could we recognize them at the recognition event? They are only having 25 years or higher at the event, but could look at it. But, could do a letter from administration thanking them for their service, and include the chair.

4. Update on performance program compliance rate: We are requesting an update on the rate of professional faculty with up-to-date performance programs.

Going from 60% and now at 88% compliance. There are 103 still outstanding. Has the quality of the performance programs improved? Some are perhaps a little better and some are just sending in a job description. On April 30, there will be a joint labor-management presentation on them. Pres. Rodriguez emphasized the importance of 100% compliance and it needs to happen on a regular basis. New deadline for those that didn't get them in is May 3.

5. Teaching guidelines for part-timers. In June of 2017 the Chapter and Management were in conversation regarding guidelines for the evaluation of teaching of part-time academic faculty. My recollection of those conversations is that they were productive, and that the Chapter was awaiting a revised draft from the CAS Dean's office. In the interest of continuing this conversation, we are requesting an update on the status of that document.

-Hedberg received the latest version and will forward it later today. The proposal is based on peer-review and review of course materials. History was that it was based on SIRF scores only. An evaluator mechanism was put in place to put into place a promotion process in the contingent track; this will also be relevant for discretionary awards.

6. Publicizing of SIRF scores: We have recently learned that the University is moving ahead with a 2012 recommendation from the Course Assessment Advisory Committee (CAAC) to publicize the 'Course Overall' and 'Instructor Overall' SIRF scores. We are requesting information on when, and how, this information will be publicized to students. We would also like to, once again, state the Chapter's opposition to policies, such as

this, that elevate the importance and impact of SIRF scores. This opposition is driven by several factors.

- a. Board of Trustees Policy (Article XII) states that the evaluation of academic faculty teaching should be based on multiple sources; University Senate Policy [see attachment 2, pp. 5-8] takes this further, stating that the SIRF scores should hold relatively little weight in these evaluations. Publicizing SIRF scores, elevates their impact which, in our view, is contrary to the letter and spirit of these policies.
- b. The Chapter's own review of the data used by the CAAC to make its recommendations shows that students' expected grade, the response rate to the SIRF and, most troubling, instructor gender, have significant impacts on SIRF scores [attachment 2, pp. 3-5].
- c. Since the CAAC completed its work, several high-quality academic studies have been published that not only undermine the claim that student evaluations measure the quality of a course or an instructor but, more troubling, show how such evaluations are systematically biased against women faculty. This is nicely summarized in a recent review written for AAUP [attachment 3]

We trust that this new policy is motivated by a desire to ensure that students are receiving the highest quality education and we share this commitment. At the same time we want to reassert that the SIRF tool does not bring us closer to that end and subjects many of our members to an evaluative process, vital for their professional success, that is discriminatory.

--Management reports that response rates have gone up (using class time to do them on their phones) and so more are being published. It doesn't just pop up, you have to search for it. Advisory committee believes that it is just one of several means of measuring teaching value. The major issue with SIRFs is whether it is an appropriate way to measure teaching effectiveness and whether it hurts faculty. There is a policy in place that requires that each unit establish a protocol on peer review of teaching. SIRFs highlight the outliers. For these faculty (e.g. WCI and EOP), there are steps to take in terms of making this a better process for contingents.

Additional Items in Conversation with President Rodriguez:

-Pres. Rodriguez wanted to highlight a few things. The university address spoke to the progress and challenges the university is facing. It was well attended, but how we can get more faculty to participate? Reviewing the UUP newsletter, Rodriguez noted that there may be a bit of disconnect in terms of what is happening at the university.

-Re: Adjunct salaries and promotions, the university is actually trying to make more progress on this. There is a lot of conversation between Pres. Rodriguez's office and Academic Affairs. They are still taking this issue very seriously. There needs to be more conversations about this.

-We are talking to membership about the DSI process and we will be bringing a set of proposals about how we'd like to see the process carried out. We would like to see the money fund part of those career ladders for contingents. We thought it was dead in the water but happy to see that this is still on the agenda.

-Management reports that the Workforce planning committee has been good and has helped to get the university out of the red. So, now they can address things like contingent salaries and graduate stipends.

-New Provost coming in on Aug. 1. She's spending some time here now too.

-Re: the university address, transcript or highlights could be made available. Talked about recording it or live streaming it and the pros and cons.

-Labor reports seeing some gains in terms of transparency, but staff still see it as another strategic planning process, or that the outcome has been pre-determined. The StAR process netted 88 proposals. They were able to fund 17 proposals. Pres. Rodriguez believes that this means that there are a lot of groups out there that are interested in the process. Management must continue to build trust.

-Management has established metrics to measure progress with the reduction of the budget deficit. If it's working they will continue, and if not, then they won't. Pres. Rodriguez is here to build bridges and trust. As issues come up, he would like us to let him know.

[Pres. Rodriguez had to leave the meeting at this point.]

-Hedberg passed out a policy concerning international travel – study abroad, research, etc. Came about because a faculty who was in Africa and got a spider bite. It exposed weaknesses in the system. Need to be clear on university's obligations. Harvey Charles has researched this. Policy would require everyone to register their plans prior to going and to confirm adequate health coverage. Developing a committee to review it, kind of like IRB. If somebody disobeys international travel committee (decides to go on their own), could they be subject to discipline. This question needs to be addressed. Also, what will the waiver language actually say re: workers' compensation. What about UUP health insurance coverage?

-Policy on faculty authored textbooks. How would this be administered? How do you re-direct the royalties? The financial piece may conflict with the minimum amount of money in the NYS conflict of interest policy and may be a negligible amount.

-Hedberg shared info about annual reviews of faculty. Came up with a form. Currently, the instrument is the FAR and they will continue to try and make it better. Anything else faculty would like to add should also be invited and solicited. Modeled after professional

faculty form, using satisfactory/unsatisfactory. If unsatisfactory, there is a meeting scheduled to discuss and create a remedy plan. What problem is this a solution for? What is the impetus? We can say that faculty are reviewed annually and we have some documentation of it. It's a best practice feature of high impact organizations. Could this be done for part-time faculty (in conjunction with what we talked about earlier)? Provides feedback between the chair and supervisor and identify areas of improvement. Part-time faculty should be filling out the FAR. The suggestion was made that this be a post-tenure process. We will have to provide training for chairs and deans. Have they met departmental norms? Could the language be changed so that it is not evaluative, i.e. meets expectations, full professional obligations. Have they done it well versus what have they done? Inventory vs. evaluation. Is it a professional obligation review? If it raises negotiability issues then this has to go back to statewide. Will give people the opportunity to remedy if they are not meeting the obligation. How can we avoid getting entangled in post-tenure review and take negative employment action? It would be good for the principle to be addressed in the materials (can't be used to force people to teach more). Training piece will be key. UUP is concerned about any kind of post-tenure review and its uses.

Attachment #1:

Resolution of Support for GSEU Fees Campaign

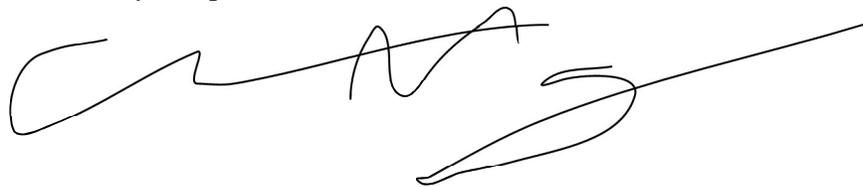
Whereas in their work as instructors, graders, researchers, and staff, graduate student employees are essential to the life and work of the university; and,

Whereas Graduate Assistants and Teaching Assistants across the SUNY system earn an average assistantship of \$15,000 and are required to pay back to the University, on average, between 14%-25% of this amount in the form of University fees; and,

Whereas as educators and University professionals, UUP is committed to supporting graduate student learning and professionalization. As Union members, we recognize that this includes being committed to the quality of life of graduate students.

Resolved that the Albany Chapter of UUP believes that University fees pose an excessive burden to graduate student employees as they fight for a living wage and offers its full support of the Graduate Student Employees Union campaign to abolish fees for graduate student employees at the University at Albany (SUNY).

Resolution adopted by the Executive Committee of
the Albany Chapter of UUP, March 25, 2019.

A handwritten signature in black ink, appearing to read 'Aaron Major', written over a horizontal line.

Aaron Major
President, UUP Albany Chapter

Attachment #2:

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 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.

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

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 analysis² 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.

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.

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

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

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 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).

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.

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

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/

Attachment #3:

Student Evaluations of Teaching are Not Valid

It is time to stop using SET scores in personnel decisions.

By John W. Lawrence

<https://www.aaup.org/article/student-evaluations-teaching-are-not-valid#.XLhI9i0pDgk>

In a review of the literature on student evaluations of teaching (SET), Philip B. Stark and Richard Freishtat—of the University of California, Berkeley, statistics department and the Center for Teaching and Learning, respectively—concluded, “The common practice of relying on averages of student teaching evaluation scores as the primary measure of teaching effectiveness for promotion and tenure decisions should be abandoned for substantive and statistical reasons: There is strong evidence that student responses to questions of ‘effectiveness’ do not measure teaching effectiveness.” This is a startling conclusion, given that SET scores are the primary measure that many colleges and universities use to evaluate professors’ teaching. But a preponderance of evidence suggests that average SET scores are not valid measures of teaching effectiveness.

There are many statistical problems with SET scores. The response rate of student evaluations is often low. There is no reason to assume that the response pattern of those who do not complete the surveys would be similar to the pattern of those who do complete them. Some colleges assume that a low response rate is the professor’s fault; however, no basis exists for this assumption. Also, average SET scores in small classes will be more greatly influenced by outliers, luck, and error. SET scores are ordinal categorical variables in which participants make ratings that can be from poor (one) to great (seven). Stark and Freishtat point out that SET score numbers are labels, not values. We cannot assume the difference between one and two is the same as the difference between five and six. It does not make statistical sense to average categorical variables.

Even if SET score averages were statistically meaningful, it is impossible to compare them with other scores, such as the departmental average, without knowing the distribution of scores. For example, in baseball, if you don’t know the distribution of batting averages, you can’t know whether the difference between a .270 and .300 batting average is meaningful. Also, it makes no sense to compare SET scores of very different classes, such as a small physics course and a large lecture class on Shakespeare and hip-hop.

More problematic are the substantive concerns. SET scores are a poor measure of teaching effectiveness. They are correlated with many variables unrelated to teaching effectiveness, including the student’s grade expectation and enjoyment of the class; the instructor’s gender, race, age, and physical attractiveness; and the weather the day the survey is completed. In a 2016 study by economist Anne Boring and statisticians Kelli

Ottoboni and Philip Stark, students in both France and the United States rated online professors more positively when they thought the instructor was a male. The researchers concluded that “SET are more sensitive to students’ gender bias and grade expectations than they are to teaching effectiveness.” In 2007, psychologists Robert Youmans and Benjamin Jee found that giving students chocolate before they completed teaching evaluations improved SET scores.

A 2016 study by two other psychologists, Nate Kornell and Hannah Hausman, reviewed the literature on the relationship between course SET scores and course performance. Course performance was measured by the same outcome measure (a final test) being given to multiple sections of the same course. In six reviewed meta-analyses, SET scores account for between 0 and 18 percent of the variance of student performance. The researchers also reviewed two rigorous studies with random assignment of students that tested whether SET scores predicted performance in a subsequent class—for example, do SET scores in Calculus 1 predict grades in Calculus 2? In both studies, student performance in the second class was negatively correlated with SET scores in the first class. Thus, students in classes with professors who received relatively low SET scores in the first semester tended to perform better in the second class. Kornell and Hausman posited that one possible explanation for these findings is that skilled teachers are able to achieve an optimum level of difficulty in their course to facilitate long-term learning. Students like it less but learn more.

Psychologist Wolfgang Stroebe has argued that reliance on SET scores for evaluating teaching may contribute, paradoxically, to a culture of less rigorous education. He reviewed evidence that students tend to rate more lenient professors more favorably. Moreover, students are more likely to take courses that they perceive as being less demanding and from which they anticipate earning a high grade. Thus, professors are rewarded for being less demanding and more lenient graders both by receiving favorable SET ratings and by enjoying higher student enrollment in their courses. Stroebe reviewed evidence that significant grade inflation over the last several decades has coincided with universities increasingly relying on average SET scores to make personnel decisions. This grade inflation has been greater at private colleges and universities, which often emphasize “customer satisfaction” more than public institutions. In addition, the amount of time students dedicate to studying has fallen, as have gains in critical-thinking skills resulting from college attendance.

If SET scores are such poor measures of teaching effectiveness and provide incentives for leniency, why do colleges and universities continue to use them? First, they are relatively easy to administer and inexpensive. Second, because they result in numerical scores, they have the appearance of being “objective.” Third, the neoliberal zeitgeist emphasizes the importance of measuring individual performance instead of working as a community to address challenges such as improving teaching. And fourth, SET scores are part of a larger problem in higher education in which corporate administrators use assessment and credentialing procedures to exert control over faculty and students. The

validity of the “assessments” is assumed and of secondary importance. In reality, the valid assessment of multifaceted phenomena such as teaching effectiveness is complex and requires scientifically rigorous investigations.

So, if SET scores are not measures of teacher effectiveness, how should colleges and universities evaluate teaching? Stark and Freishtat assert that measuring both teaching effectiveness and learning is complex and that we cannot do so reliably and routinely without applying rigorous experimental methods to all courses. They suggest that we focus on evaluating teaching. This requires evaluating the materials that professors create for classes and doing regular teaching observations. Yet, as a former department chair, I have read hundreds of peer teaching observations, and my impression is that the interrater reliability of peer observations is itself not particularly high.

Given the complexity of measuring good teaching, colleges and universities need to engage in this task with humility. It may be more fruitful for the institution to approach teaching observations as a strategy for facilitating collegial dialogue on striving for high quality teaching and experimentation. The emphasis needs to be on mentoring, not potentially punitive evaluations. Moreover, in regard to improving student performance, institutional changes are likely more effective than focusing on individual professor performance. Particularly for minority and first-generation students, small class sizes and increased opportunities to interact with the professors and peers can improve a variety of outcomes.

John W. Lawrence teaches psychology at the City University of New York College of Staten Island. He is also a grievance counselor for the union representing CUNY faculty and staff, the Professional Staff Congress. His email address is john.lawrence@csi.cuny.edu.