NEW YORK UNIVERSITY

Full-Time Continuing Contract
Faculty Senators Council

411 Lafayette Street, Room 327
New York, NY 10003

P: 212 998 2230

c-fsc@nyu.edu

Date: November 14, 2018

Memo to: Katherine Fleming, Provost

From: Mary Killilea
Chairperson, C-Faculty Senators Council
A/Y 2018-2019

Subject: C-Faculty Senators Council Report and Recommendations on Student Evaluations

Please see the attached report and recommendations that was prepared by the C-FSC Educational Policies and Faculty/Student Relations Committee and endorsed by the full C-FSC on May 11, 2017. Minor revisions were made over the summer of 2017 and the report should have been sent to you last fall but was over looked. I apologize for the delay but still think the report is important and relevant.

cc: Gigi Dopico, Vice Provost for Undergraduate Academic Affairs
Carol Morrow, Vice Provost
Ellen Schall, Senior Presidential Fellow

C-FSC Steering Committee Members:
Larry Slater, C-FSC Vice Chairperson
Lauren Davis
Leila Jahangiri
Tommy Lee
Beverly Watkins
Ethan Youngerman

Scott Illingworth, C-FSC Educational Policies and Faculty/Student Relations Committee Chair

Wen Ling, T-FSC Chairperson

Continuing Contract Faculty Senators Council Recommendations On Student Evaluations:

Report Prepared by the C-FSC Educational Policies and Faculty/Student Relations Committee
Committee Members:

Ben Stewart, Chair, Faculty of Arts and Science
Spiros Frangos, School of Medicine
Neal Herman, College of Dentistry
Brian Mooney, School of Professional Studies
Deborah Smith, School of Professional Studies

Date: 8/17/2017

Introduction

This report has been written to begin a conversation about the revision and use of student evaluations of teaching (SETs) at NYU. The Continuing Faculty Senators Council (C-FSC) seeks to ensure that SETs are used primarily in the interests of identifying and cultivating good teaching. We are especially concerned with this issue given that a number of studies have raised concerns about various kinds of bias within teaching evaluations. Moreover, within the scholarship on SETs, there is an ongoing debate about problems with global measures of teaching (where global measures refer to general questions about an instructor or a class). Given that evaluations are regularly used in the assessment of continuing faculty, these issues are of special concern to our constituents.

We propose four sets of recommendations, two of which request collaboration with the Provost's office and two of which make more general recommendations. The recommendations addressed to the Provost's office focus on the revision and assessment of NYU SETs. Our other recommendations address issues related to the use of SETs for assessment.

Requests for Collaboration with the Provost's office:

I. Recommendations for the Revision and Assessment of NYU's SET system

1. We recommend that NYU's Office of Institutional Research examine the extent to which the gender, age, and race of teachers and students affects student evaluations.
2. We recommend the revision or deletion of the evaluation's global measures (See Appendix A, questions 1 and 2: "Overall evaluation of the instructor(s)," and "Overall evaluation of the course"). If questions 1 and 2 are revised, we recommend that they be replaced with more specific questions.
3. While those global measures remain in place, we recommend that NYU's Office of Institutional Research examine the extent to which students' responses to those

questions correlate with or diverge from the their answers to the evaluation's more specific questions.

II. Recommendations for the Customization of Evaluation Questions at the School Level:

1. Continuing/contract faculty should be represented on the School-level committees that develop and revise evaluation questions for the customizable portion of the NYU evaluation. These faculty members should be elected faculty representatives (either from School- or University-level bodies).
2. Quantitative questions should err on the side of asking about specific rather than general aspects of the class and the teaching (see footnote 1).
3. Qualitative (short answer) questions should be framed so as to encourage specificity in students' answers. Consider the qualitative question on the customized portion of CAS evaluation: "Describe the best thing about the course/instructor that was effective in helping you learn." We see that question as a good model in that it encourages students to focus on a concrete skill that they took from the class.

Recommendations for Using SETs as Assessment Tools

III. Recommendations for Faculty Who Are Up for Reappointment and/or Promotion.

1. Faculty members should carefully review the evaluations and consider whether the responses suggest worthy changes in pedagogy.
2. We recommend that faculty consider evaluations in the context of longer-term patterns of response. While we know that the responses in any one class or for any one semester are not necessarily indicative of much, patterns that persist over time and across different courses are stronger indicators of areas for improvement.
3. Departments and programs should establish peer mentoring or other forms of peer support to cultivate faculty reflection on their evaluations, particularly in relation to reappointment. We recommend that these peer interactions include discussion of the advantages and disadvantages of including reflections on evaluations in reappointment and promotion documents (how much or how little to include, how best to frame those reflections, etc.).

IV. Recommendations Related to Administrative Assessment.

1. SETs (despite their name) should primarily be used to assess aspects of the curriculum rather than to evaluate the performance of individual teachers.
2. Comparison of SETs among programs, departments, or schools should be avoided.
3. Evaluations of faculty should avoid the use of quantitative data from student evaluations. School-based Reappointment and Promotion (R&P) recommendations should be re-examined to address this recommendation. At most, such numbers should be used to sensitize the reading of qualitative data.
4. To the extent that quantitative data are considered, they should only be examined longitudinally. A given class or semester may be an outlier. At best, quantitative may suggest a trend in teaching quality over a period of time.
5. For a given class, R&P recommendations should attend carefully to the statistical validity of the sample (both in terms of the size of the class and the response rate).
6. Student evaluations are best examined holistically. The faculty member familiar with the course and with the students enrolled in it can often put student responses in context. Deans and administrators should be discouraged from looking at the student responses without discussing them with the faculty member; faculty should have an opportunity to explain the context for students' responses. The faculty member is likely to have valuable insight into which of the student narrative responses are worthy of consideration and which ones are either incorrect or false claims, or are possibly motivated by some other personal reason.
7. Beyond administrators, R&P committees, and those involved in program assessment, qualitative data should remain confidential.

Background for C-FSC Recommendations on Evaluations

We are concerned that SETs may not always work in the interests of the evaluated faculty members or to the institutions to which they belong. A number of recent studies (Anderson and Miller 1997; Basow 1995; Boring, Ottoboni, and Stark 2016; Cramer and Alexitch 2000; Stark and Freishtat 2014) have questioned the validity of SETs. These studies suggest that SETs exhibit a range of student biases with respect to gender, ethnicity, and age. Some of these distortions are complexly layered. For instance, Basow (1995) finds evidence that the gender of the student is significant, as is the academic division in which evaluations take place:

> male faculty are perceived and evaluated similarly by their male and female students, whereas female faculty tend to be evaluated differently, depending on the divisional affiliation of the course. Female faculty tend to be rated highly by their female students, especially in the humanities, but less positively by their male students, especially in the social sciences (664).

Such evidence led us to our first recommendation, that NYU's Office of Institutional Research should "examine the extent to which the gender, age, and race of teachers and students has effects on student evaluations."

Additionally, we are concerned with the general character of the first two questions on NYU's evaluation, which led us to our second recommendation, namely, the "revision or deletion of the evaluation's global measures" (See Appendix A, questions 1 and 2). During the period in which those questions remain on the evaluation, we also hope that the Office of Institutional Research will "examine whether students' responses to the global questions correlate with or diverge from the their answers to the evaluation's more specific questions."

On the issue of "global measures," Gravestock and Gregor-Greenleaf (2008) argue that "Most attempts to identify particular characteristics of effective teaching stem from a belief that teaching should be measured according to multiple aspects or categories of teaching activity" (31). However, they also note lingering questions about how to present overall measures of teaching—should measures of multiple dimensions be averaged, or is there value to questions that call for overall, global evaluations of teaching?

While there is no consensus on the value or dangers of global questions, Gravestock and Gregor-Greenleaf end their discussion of that issue with a caution about questions such as those "that ask students if they would recommend the course to others" (32) and they also describe a change that the University of Minnesota made to their student evaluation. In 2007 (109),

> the University of Minnesota decided to eliminate its global question, "How would you rate the instructor's overall teaching ability?" The committee charged with revising the instrument argued that this item was too often the only score evaluated in summative teaching assessment, that students have difficulty responding to the question, that the item is not diagnostic and that global questions such as these do not correlate with ratings on questions that review specific teaching characteristics. (32)

Issues with the evaluation instrument may be further complicated as a result of biases that emerge out of the relations between teachers and students. For example, Wolfgang Stroebe (2016) outlines a possible cause of such bias. Stroebe theorizes that, because student evaluations are such important "determinants of academic personnel decisions" (801), teachers may exchange leniency (in the form of inflated grades) for higher student evaluations. In support of his claim, he primarily cites a number of psychological studies that suggest students' inclination to give teachers lower evaluation scores when they receive lower-than-expected grades. To the extent that teachers make their classes more lenient as a result of their concerns about evaluation scores, high evaluation scores "reflect[] a bias rather than teaching effectiveness" (800).

Although it's clear that the linkage between teaching evaluations and reappointment creates incentives for teachers to give higher grades, we are not so certain that grade inflation is a problem among NYU continuing faculty members. In fact, we are concerned that NYU's continuing faculty may be giving grades that are *lower* than NYU students' average grades. Consider, for instance, the percentage of A-range grades (A and A-) that the Expository Writing Program (EWP) gives relative to CAS Humanities classes (including EWP) and to CAS in general:

Percentages of A-range grades:

	EWP	CAS Humanities	All of CAS
F 2014	29%	52%	48%
S 2015	45%	60%	52%
F 2015	36%	56%	49%
S 2016	44%	61%	52%

EWP's faculty includes one tenured faculty member and approximately 105 continuing faculty. Note that, whereas the Fall numbers include the grades of all of the students EWP teaches, the Spring percentages do not include the grades of EWP students in Tisch and Steinhardt (whose Spring classes run under those schools' course codes). The higher number of A-range grades that EWP faculty give in the Spring may partly be due to those students having had the experience of a semester of college before they take the course.

To the extent that this is so, it begins to clarify why EWP gives a lower number of A-range grades—i.e., it makes sense that students would receive lower grades when they are new to college as compared to the grades they receive later. Nonetheless, this factor also highlights the difficulty of the teaching role that continuing faculty are often in—namely, that of introducing students to the conventions and expectations of the university, and of doing so in situations where the students lack a clear sense of why they need to understand those conventions or why they should attend to those expectations.

For the teachers in that position, evaluation numbers may not always provide a clear signal of teaching quality: in some cases, low evaluations may signal rigorous teaching; in other cases high evaluations may signal an avoidance of the difficulties of helping students to take on scholarly conventions, especially when it comes to gaining the knowledge learning practices that will benefit them over the long term. For instance, the results of Carrell and West's (2010) seven-year-long study, which looked at multiple years of evaluations from 10,534 students, suggest "that evaluations reward professors who increase achievement in the contemporaneous course being taught, not those who increase deep learning" (430). As *continuing* faculty, we want to encourage the kinds of teaching that have larger payoffs down the road, not only for our students, but also for the teachers who will interact and engage with those students in the future. We are concerned that an overvaluing of SETs—and, separately, a *perception* among faculty that the SETs are overvalued—could disincentivize precisely the kind of teaching and learning that rigorous evaluation of faculty is meant to ensure.

Given that NYU is currently in the process of transitioning to a University-wide evaluation system, it's an especially important time to establish procedures around that system's revision, assessment, and use for purposes other than providing feedback to individual teachers. This situation motivates those of our recommendations that extend beyond the evaluation instrument itself: those that call for continuing faculty involvement in the development and revision of questions; those that suggest strategies for teachers to engage with their evaluations; and finally, those that offer protocols for the use of evaluations in assessment, especially in reappointment and promotion decisions.

Works Cited

Anderson, K., & Miller, E.D. "Gender and Student Evaluations of Teaching." *PS: Political Science and Politics*, 30.2 (1997): 216-219.

Basow, S.A. "Student evaluations of college professors: When Gender Matters." *Journal of Educational Psychology*, 87.4 (1995): 656-665.

Boring, Anne, Kellie Ottoboni, and Philip B. Stark. "Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness." *ScienceOpen Research* (2016): 1-11.

Carrell, Scott E., and James E. West. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors." Journal of Political Economy 118.3 (2010): 409-432.

Cramer, K.M. & Alexitch, L.R. "Student Evaluations of College Professors: Identifying Sources of Bias." *Canadian Journal of Higher Education*, 30.2 (2000): 143-64.

Gravestock, Pamela, and Emily Gregor-Greenleaf. *Student Course Evaluations: Research, Models and Trends*. Toronto: Higher Education Quality Council of Ontario, 2008.

Stark, Philip B., and Richard Freishtat. "An evaluation of course evaluations." *ScienceOpen Research* 9 (2014): 1-26.

Stroebe, Wolfgang. "Why Good Teaching Evaluations May Reward Bad Teaching: On Grade Inflation and Other Unintended Consequences of Student Evaluations." *Perspectives on Psychological Science* 11, no. 6 (2016): 800-816.

Appendix A: Evaluations from CAS

CAS Evaluation Questions

General Questions (University: all on a 5-point, Likert scale)

1	Overall evaluation of the instructor(s)
2	Overall evaluation of the course.
3	The instructor(s) provided an environment that was conducive to learning.
4	The instructor(s) provided helpful feedback on assessed class components (e.g., exams, papers).
5	The course objectives were clearly stated.
6	The course was well organized.
7	The course was intellectually stimulating.

CAS Custom Questions (8-17 on a 5-point, Likert scale; 18 is a qualitative question)

Questions about the Course

8	The course was effective at helping me learn.
9	The classes were informative.
10	The course was challenging.
11	The course increased my knowledge of the subject.

Questions about the Instructor

12	The instructor was effective at helping me learn.
13	The instructor encouraged student participation.
14	The instructor was effective at facilitating class discussion.
15	The instructor was open to students' questions and multiple points of view.
16	The instructor was accessible to students (e.g., via e-mail and office hours).
17	The instructor created an environment that promoted the success of students with diverse backgrounds.

Qualitative Question:

18	Describe the best thing about the course/instructor that was effective in helping you learn.