SENATE COMMITTEE ON FACULTY AFFAIRS

Student Rating of Teaching Effectiveness (SRTE) Evaluations:
Effective Use of SRTE Data

(Informational)

Introduction

Senate Chair James Strauss asked the Faculty Affairs Committee to sponsor a report from Senator Angela Linse on the Student Ratings of Teaching Effectiveness (SRTE). This report focuses on using student ratings data in the faculty evaluation process and is based on Senator Linse's original work (Linse, in press), with additions specific to Penn State and the SRTEs.

The purpose of this report is to provide guidance about some of the most common misuses of student ratings data in the faculty evaluation process, and to set forth guidelines for best practices in the use and evaluation of SRTEs. The report includes a brief overview of student ratings, including a description of what student ratings are and are not (or, in other words, what they do and do not do). The most important sections are the two sets of guidelines developed for two distinct target audiences responsible for faculty evaluation—faculty serving on review committees and academic administrators. The two sets of numbered guidelines are written so that they may be distributed separately.

Faculty have a host of reasonable concerns about student ratings. Most of these concerns have been the subject of more than 80 years or research, which has been published in a vast body of peer-reviewed literature. This research generally refutes common misperceptions, but the literature is not widely known or readily accessed by faculty or administrators.

Faculty in evaluative roles are rarely, if ever, provided guidance on interpreting other faculty members' student ratings even though faculty regularly rotate onto review committees and move into new administrative roles. Without research-based guidance, these faculty end up relying on their own experiences, biases, and possibly erroneous information to the evaluation process. New administrators will eventually see a wide range of student ratings and typically develop a broader understanding of the variability across courses and individuals. However, faculty on review committees may only see the ratings for a few faculty per year, which could encourage over-interpretation of the results.

Overview

Student ratings instruments have been around since the 1920s (Marsh, 1987; Remmers, 1933; Remmers & Brandenburg, 1927). The term "student ratings" refers to surveys directly administered by colleges and universities to enrolled students under controlled circumstances, typically near the end of an academic term. Other common names for these surveys include student evaluations of teaching (SETs), student ratings of instruction (SRIs), teaching evaluations, and course evaluations.

Student ratings are nearly ubiquitous in U.S. higher education and the practice has become more common in other countries in the past few decades (Berk, 2005; Seldin, 1999; Miller & Seldin, 2014). Student ratings serve as a source of feedback for instructional improvement, but they are also widely used in personnel decisions such as annual reviews, merit raises, promotion and tenure, post-tenure review, and for hiring and re-appointment of "tenure exempt" faculty.[1]

Penn State's Faculty Senate created the Student Ratings of Teaching Effectiveness (SRTE) instrument to improve the consistency of student ratings across the university for the purposes of promotion, tenure, and annual review (see University Faculty Senate Record, Vol. 19, No. 6, April 30, 1985). Penn State is unlike other institutions where student ratings instruments originally created for developmental purposes were co-opted for evaluative purposes. Approval of the April 30, 1985 Advisory and Consultative report established the *Statement of Practices for the Evaluation of Teaching Effectiveness for Promotion and Tenure*, as a set of guidelines maintained separately from the *Promotion and Tenure Procedures and Regulations* (HR21) and the associated *Administrative Guidelines for HR23*. While separately maintained, the Statement of Practices appears as Appendix A in the Administrative Guidelines (http://vpaa.psu.edu/files/2016/09/p_and_t_-guidelines-2i76gdt.pdf). The 1985 senate legislation explicitly repealed all previous senate legislation related to the evaluation of teaching. By fall semester 1987, faculty across the university had provided the items now included in the common pool of 177 questions (see http://srte.psu.edu/SRTE_Items/).

The challenge of appropriate use of student ratings data will be with us as long as we continue to use them. When student ratings are used in personnel decisions, it is critical that they be used appropriately, and in ways consistent with the recommendations of experts in student ratings research (McKeachie, 1997; Theall & Franklin, 2001). Excellent reviews and summaries of the research literature include publications by Benton & Cashin (2011), Benton & Li (2015), Berk (2005, 2013), Cashin (1999, 2003). Faculty interested in improving their understanding of the history and leading researchers in the student ratings field should consider reading Feldman (1976, 1989, 1992, 1993, 2007), Franklin and Theall (Franklin, 2001, Franklin & Theall, 1991, 1994; Theall & Franklin, 1990, 2000, 2001), Hativa (2013b), Marsh (1980, 1982a, 1982b, 1984, 1987, 2007; Marsh & Dunkin, 1992; Marsh & Roche, 1997), McKeachie (1979, 1990, 1997) and Ory (2001; Ory & Ryan, 2001; Ory et al., 1980). Arreola (2007), Berk (2006), Braskamp et al. (1984), Cashin (1996), and Hativa (2013a) provide guidance on how to create a valid and reliable faculty evaluation system.

What Student Ratings Are and Are Not

Before advancing to the principal sections of this article (Guidelines for Faculty and Questions Asked by Administrators), it is important to clarify what student ratings are and are

1. I prefer to use a positive term, "tenure exempt," to describe a class of faculty that has long been the majority in most U.S. colleges and universities, rather than the more typical terms "non-tenure-line" and "adjunct" faculty. The latter terms marginalize these faculty because they describe what they are *not*, emphasize difference, and highlight a lack of status.

not. The purpose of the items below is to clarify that "students' classroom experiences," "student learning," and "teaching quality" are not synonymous.

- **Student ratings are student perception data**
  Student ratings instruments are used to gather the collective views of a group of students about their experience in a course taught by a particular faculty member (Arreola, 2007; Abrami, 2001; Hativa, 2013a). Data are typically collected systematically from enrolled students who have experienced the learning environment created by the faculty member. Students can provide a unique perspective on how the instruction and the course requirements affected them and they have had many opportunities to observe the teacher in class and interacting with students—these are perspectives that other faculty cannot provide. Because students provide these unique perspectives on the teaching and learning environment, their collective perceptions can serve as one source of data for teaching improvement and for personnel decisions, as well as for research (McKeachie, 1990, p. 194).

- **Student ratings are not measures of student learning**
  Student ratings data were never intended to serve as a proxy for learning, but instead as a method of gathering student perception data. Even though assessment of student learning is rightly in the domain of the faculty member, that student ratings do not measure learning is a common criticism. Conflation of gathering student perceptions and assessing student learning may stem from research that has demonstrated a low to moderate positive correlation between students' ratings and their grades or expected grades (Abrami et al., 1980; Abrami, 2001; Benton & Li, 2015; Eiszler, 2002; Feldman, 1976; Greenwald & Gilmore, 1997; Stumpf & Freedman, 1979). Correlation is to be expected.

  When students learn and subsequently receive high grades for demonstrating their learning, some might understandably assign credit for their learning to the faculty member. That said, a simple correlation between grades and student ratings does not demonstrate causality. We cannot assume that if a faculty member gives high grades, the faculty member will receive high ratings. At Penn State, faculty members whose students learn also tend to receive higher ratings from their students—and deservedly so.

- **Student ratings are not faculty evaluations**
  Students are just one source of data used in the evaluation of faculty. Students are *producers of data* and faculty and administrators are the *interpreters of that data*, whether the data are used for improvement or evaluative purposes. Faculty and administrators are responsible for deciding how to use and interpret the ratings data provided by students. Unfortunately, the names that colleges and universities assign to their ratings instruments (Student Evaluations of Teaching, Course Evaluations) probably contribute to the view that student ratings constitute an instance of evaluation, rather than an instance of data collection.

- **Student Ratings are here to stay**
  Given the utility of student ratings in academic personnel decisions, student ratings are unlikely to be eliminated any time soon (Benton & Cashin, 2011; Kulik, 2001; Franklin, 2001). Furthermore, most faculty agree that students' views should not be entirely ignored (Berk, 2006). As such, how these data are interpreted and (mis)used is important (McKeachie, 1997).

Guidelines for Faculty Who Use Student Ratings Data to Evaluate Other Faculty

1. *Student ratings should be only one of multiple measures of teaching.*

Student ratings proponents and researchers unanimously recommend personnel decisions be based on more than just the faculty member's student ratings (Arreola, 2007; Benton & Cashin, 2011; Benton & Li, 2015; Berk, 2013; Cashin, 1996, 1999, 2003; Hativa, 2013a; Marsh, 1987; McKeachie, 1990, 1997; Miller & Seldin, 2014; Nulty, 2008). The most common additional sources of data about the faculty member's teaching include written student feedback, peer and administrator observations (Miller & Seldin, 2014), internal or external reviews of course materials (Chism, 2007; Miller & Seldin, 2014), and more recently, teaching portfolios (Seldin, 1999; Zubizarreta, 1999) and teaching scholarship (Berk, 2013; Miller & Seldin, 2014). While none of these additional data collection methods have been extensively examined for reliability, validity, or bias (as have student ratings), they provide important points of comparison to students' perspectives. Data collection for each of these additional data sources should be systematic rather than informal.

Penn State guidelines recommend against using SRTEs as the sole measure of teaching effectiveness. This is stated in a number of places and ways in the *Statement of Practices for the Evaluation of Teaching Effectiveness for Promotion and Tenure*. The final line in the Statement of Practices document is "Furthermore, student evaluations alone are not sufficient for either personnel decisions or for improvement of teaching" (Section I.D.4.c).

2. *In personnel decisions, a faculty member's complete history of student ratings should be considered, rather than a single composite score.*

Some academic units (departments, colleges, schools) combine a faculty member's cumulative record into a single score. Cashin (1999) recommends looking across time and courses in order to generalize about students' views of an instructor's teaching and discourages creating a single score, in part because teaching is multidimensional (Abrami, 2001; Franklin, 2001; Marsh,1984; Marsh & Dunkin, 1992) and is difficult to represent in a single score. The temptation to create a composite score may derive from the common practice of tenure and promotion committees to label each faculty member's research, teaching, and service with a single evaluation along a scale from excellent to poor. While statistical models can be used to create a composite score that weights different teaching factors (Marsh, 1987), the adjustments should be applied to all faculty. If only some faculty have their ratings reduced to a single number, they are being evaluated under different circumstances than their peers and may be at a significant disadvantage. Furthermore, evaluators can be assured that the results are reliable

when they see similar ratings across multiple courses because “multiple classes provide more reliable results than a single class” (Benton & Cashin, 2011). Creating weighted averages or adjusted means based on perceptions about the ease or difficulty of teaching a particular type of course or teaching context should be avoided (e.g., adding a 0.2 points for teaching a course larger than 50).

Another reason to avoid reducing a faculty member’s student ratings history to a single composite score is that anomalous ratings are given the same weight as average ratings that are more common and consistent. A faculty member with a single cumulative rating may be unfairly disadvantaged relative to faculty whose entire history is visible and for whom anomalous scores can be explained and/or disregarded (see Table 1). The hypothetical faculty member represented in Table 1 would have a lower composite average for the Overall Course rating if the anomalous results were not differentiated. These anomalous results in Table 1 are explainable as the result of a low number of responses in a very small course (three respondents out of seven students), a low response rate (37%) in course D, year 4, and a possible curricular problem with another course (F).

3. Small differences in mean (average) ratings are common and not necessarily meaningful.

Student ratings are "broad brush" instruments used to gather information from a group of students, not all of whom will agree. They are not precision tools that produce a measurement that can then be compared to a known standard. Unfortunately, some faculty evaluators over-interpret small differences as indicative of a problem, a decrease in quality, or as an indication that one faculty member is materially better than another. In reality, a faculty member could teach the same course under similar conditions and in a similar way and still receive results that differ. Sources of variation include differences in the students enrolled, in student ratings respondents, and chance.

Variations of up to 0.4 points within a course are not unusual, but that variability can change depending on the number of categories in the ratings scale (Cashin, 1999; Husbands, 1997; Marsh, 1980, 1982a, 1982b). Rather than focusing on small differences in average scores that may not be meaningful (Abrami, 2001; Ory & Ryan, 2001), evaluators’ time is better spent looking for patterns and consistency within courses and across time (Pallett, 2006). Table 2 shows the same set of ratings as Table 1, but reorganized by course and in chronological order.

This alternative perspective shows that course F consistently receives low overall course ratings while the faculty member receives high overall instructor ratings, which may indicate a curricular problem rather than an instructional issue. Given that review committees typically do not have access to the ratings of all faculty that teach a single course, reviewers must rely on contextual commentary provided by a department head or program chair, who may be able to confirm that the course is consistently rated low by students regardless of the faculty member. This commentary can help evaluators not attribute low ratings directly to the faculty member’s teaching.

The argument for not over-interpreting relatively small differences in average ratings is supported by research that indicates a wide variety of factors have relatively small impacts on

student ratings, but that none of these alone, or even in combination can explain extremely low ratings for a faculty member. These include class size, course level, major vs. non-major courses, elective vs. required, and discipline (Feldman, 2007; Arreola, 2007; Hativa, 2013b). Bias due to gender, race, ethnicity, or culture is addressed in the next section under the question about student bias (5).

Table 1. A hypothetical faculty member's comprehensive history of student ratings (1-7 Likert scale, with 1 the lowest and 7 the highest rating); possible anomalies are indicated in bold.

Year	Semester	Course	Enrollment	Response Rate	Overall Course	Overall Instructor
1	Fall	A	125	51%	5.72	5.26
1	Fall	A	126	49%	5.98	5.34
1	Fall	B	35	43%	5.60	5.81
1	Spring	A	73	68%	5.87	5.52
1	Spring	B	29	52%	5.73	5.96
1	Spring	B	29	47%	5.76	6.32
2	Fall	A	136	41%	6.01	5.57
2	Fall	B	38	25%	5.53	5.64
2	Fall	C	9	66%	5.23	5.74
2	Spring	A	95	56%	6.32	5.62
2	Spring	B	32	57%	5.98	6.17
2	Spring	E	19	47%	5.22	5.44
3	Fall	A	90	54%	6.21	5.89
3	Fall	B	38	61	5.86	6.56
3	Fall	C	**7**	**43%**	**2.75**	**4.42**
3	Spring	A	102	49%	6.50	5.77
3	Spring	B	32	67%	6.00	6.41
3	Spring	E	12	50%	5.51	5.50
4	Fall	A	143	45%	5.08	5.58
4	Fall	C	5	48%	5.87	6.09
4	Fall	E	17	71%	5.25	5.47
4	Fall	F	55	52%	**4.49**	5.84
4	Spring	D	27	**37%**	**4.93**	5.90
4	Spring	E	23	61%	6.23	6.69
5	Fall	C	8	75%	5.75	6.17
5	Fall	E	40	78%	5.22	5.63
5	Fall	F	65	64%	**4.44**	6.85
5	Spring	D	24	63%	5.15	6.25
5	Spring	F	40	55%	**4.25**	5.48
5	Spring	F	50	33%	**4.78**	6.00

Table 2. A hypothetical faculty member's student ratings history ordered chronologically by course (1–7 Likert scale, with 1 the lowest and 7 the highest score); possible anomalies are indicated in bold.

Year	Semester	Course	Enrollment	Response Rate	Overall Course	Overall Instructor
1	Fall	A	125	51%	5.72	5.26
1	Fall	A	126	49%	5.98	5.34
1	Spring	A	73	68%	5.87	5.52
2	Fall	A	**136**	41%	6.01	5.57
2	Spring	A	95	56%	6.32	5.62
3	Fall	A	90	54%	6.21	5.89
3	Spring	A	102	49%	6.50	5.77
4	Fall	A	**143**	45%	**5.08**	5.58
1	Fall	B	35	43%	5.60	5.81
1	Spring	B	29	52%	5.73	5.96
1	Spring	B	29	47%	5.76	6.32
2	Fall	B	38	**25%**	**5.53**	**5.64**
2	Spring	B	32	57%	5.98	6.17
3	Fall	B	38	61%	5.86	6.56
3	Spring	B	32	67%	6.00	6.41
2	Fall	C	9	67%	5.23	5.74
3	Fall	C	7	43%	**2.75**	**4.42**
4	Fall	C	5	48%	5.87	6.09
5	Fall	C	8	75%	5.75	6.17
4	Spring	D	27	**37%**	**4.93**	5.90
5	Spring	D	24	63%	5.15	6.25
2	Spring	E	19	47%	5.22	5.44
3	Spring	E	12	50%	5.51	5.50
4	Fall	E	17	71%	5.25	5.47
4	Spring	E	23	61%	6.23	6.69
5	Fall	E	**40**	78%	5.22	5.63
4	Fall	F	55	52%	**4.49**	5.84
5	Fall	F	65	64%	**4.44**	6.85
5	Spring	F	40	55%	**4.25**	5.48
5	Spring	F	50	**33%**	**4.58**	6.00

The argument for not over-interpreting relatively small differences in average ratings is supported by research that indicates a wide variety of factors have relatively small impacts on student ratings, but that none of these alone, or even in combination can explain extremely low ratings for a faculty member. These include class size, course level, major vs. non-major courses, elective vs. required, and discipline (Feldman, 2007; Arreola, 2007; Hativa, 2013b). Bias due to

gender, race, ethnicity, or culture is addressed in the next section under the question about student bias (5).

Penn State's *Statement of Practices for the Evaluation of Teaching Effectiveness for Promotion and Tenure* explicitly recommend against assigning the student ratings "a precision they do not possess" (Section I.D.4.a). A report from the Senate Committee on Faculty Affairs notes that "the spirit of the SRTE is best served by regarding SRTE results as the students' view of the candidate's teaching effectiveness in absolute terms – that Professor X (whose evaluation mean is 6.25) is a "very good teacher," without necessarily saying that Professor X is a better "very good teacher" than Professor Y (whose evaluation mean is 6.10)" (Senate Record, Vol. 22, No. 6, February 21, 1989)

4. *Treat anomalous ratings for what they are, not as representative of a faculty member's teaching.*

Look for patterns in the faculty member's scores over time or across different course types. Do they show a general improvement or a persistent and unexamined issue? *Every* faculty member, even the very best, receives an occasional low average rating (Franklin, 2001). And every faculty member will have a course that does not go well or a course with unhappy students. When reviewing other faculty members' scores, patterns of low scores are more important than occasional low scores. For example, some faculty are more comfortable teaching particular types of courses. Also look for patterns of improvement that post-date a low rating, which may provide evidence that the faculty member is making improvements.

Table 2 highlights that some of the ratings of our hypothetical faculty member do appear to be anomalous. For example, the 5.08 average rating for course A in the fall of her fourth year is inconsistent with previous ratings. This anomalous rating can be explained by a substantial increase in enrollment, which could have resulted in students viewing the course as impersonal. The rating does not necessarily indicate that the faculty member cannot teach well in large courses, but it may indicate a need to adjust in-class activities. Table 2 shows many positive trends, including that the faculty member's scores are generally consistent within and across courses and that her scores have improved over time. These patterns are more important than a few low ratings over the course of five years.

5. *Examine the distribution of scores across the entire scale, as well as the mean.*

Most student ratings scores are ordinal-, not ratio-level, so the difference between a mean of 5.9 and a 6.2 (on a 7-point scale) is not meaningful when considered from the students' perspectives. Relying solely on the mean, without examining the overall shape of the distribution and the spread of scores can provide an inaccurate picture of the students' views.

Very few faculty have a normal distribution of scores (Theall & Franklin, 1990). Student ratings distributions are typically negatively skewed (Arreola, 2007; Hativa, 2013a, 2013b), i.e., they have a long tail at the low end of the scale and the mode at the high end of the scale. This tells us that most students have positive views of their courses and instructors and the mean (average) is not the best measure of central tendency for skewed distributions. Means are more

appropriately used with normal (bell-curve) distributions. In skewed distributions, means are sensitive to (influenced by) outlier ratings; in student ratings, these outliers are almost always low scores.

In small-enrollment courses, even one or two low scores can shift the mean lower, even though those students' views are not representative of the majority of students. The median or the mode is a better measure of central tendency in skewed distributions, but only a few instruments use the median or also report the median (e.g., Student Ratings of Instruction, IDEA Center; Instructional Assessment System, University of Washington).

Any report of a mean or median should also include the distribution of scores across the scale or a bar chart of the scores. If it is not possible to include the distribution with the mean or median, there may be other ways to ensure that reviewers have this additional information. For example, some institutions provide department heads with an opportunity to provide a narrative about the faculty member's teaching, which would be a good place to mention the distribution of both scores and student comments.

6. *Evaluate each faculty member individually. Evaluations and decisions should stand alone without reference to other faculty members; avoid comparing faculty to each other or to a unit average in personnel decisions.*

Student ratings instruments are not designed to gather comparative data about faculty (Franklin, 2001). The purpose of these instruments is to get an overall sense of the students' perceptions of a single faculty member teaching a particular course (or part of a course) to a specific group of students. Ultimately, no one faculty member teaching a group of students can be assumed to have the same experience as a different faculty member, even if he/she is teaching the same group of students (McKeachie, 1979).

The faculty who are most likely to be negatively impacted by faculty-faculty comparisons are those who do not fit common stereotypes about the professoriate—typically women and faculty of color. Biases, even unconscious biases, against non-majority faculty are well known in the academy (Gutgold & Linse, 2016), especially in white-male-dominated fields such as business and the STEM (Science, Technology, Engineering & Math) disciplines (National Academies, 2006; Street et al., 1996). However, such bias can also negatively impact any faculty member who is seen as different by students and faculty evaluators.

If personnel decisions are made by comparing faculty to each other, but only in some units, the faculty of those units are at a disadvantage relative to other faculty in units that do not compare faculty to each other. Faculty evaluators and administrators are the only people with the power to stop this practice.

Unit means are not an appropriate cutoff or standard of comparison because there will always be some faculty members who are, by definition, "below the mean." This is particularly problematic in units with many excellent teachers. Consider the case of a department with a mean of 6.0 on a 7-point scale. If the departmental mean is the "standard" of comparison, then

faculty who have a mean of 5.5 or even a 5.9 will be labeled as "below the mean" despite being rated by students as very good teachers (Arreola, 2007).

7. *Focus on the most common ratings and comments rather than emphasizing one or a few outlier ratings or comments.*

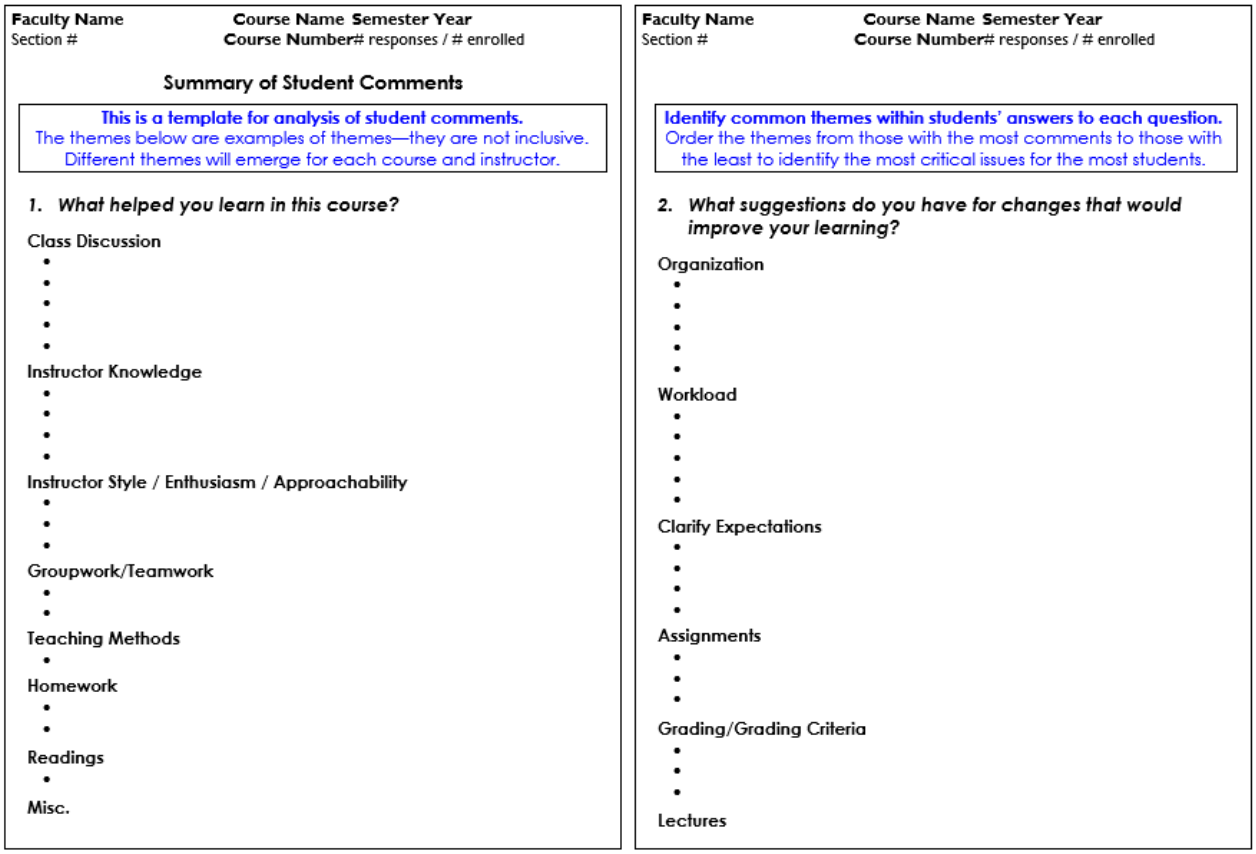
Student ratings instruments are designed to reflect the collective views of a sample of students. They are best at capturing the modal perceptions of respondents, but they are not the best instruments for capturing rare views, i.e., the views of students represented by the tail of the distribution. While students with outlier views are not unimportant, they should not be given more weight than the views of most students. This is particularly crucial when evaluating the ratings of non-majority faculty because we often see students with biased views represented in the tails of the distribution.

Many student ratings instruments are accompanied by additional questions that request written feedback from students. A variety of research indicates that written comments are highly correlated with student ratings (Berk, 2005; Braskamp et al., 1981; Marincovich, 1999; Ory et al., 1980). Too often, faculty and administrators seem to focus their attention on rare ratings or comments, possibly because the written comments are typically the most vehement or the most negative (Franklin, 2001; Franklin & Berman, 1998). It is neither appropriate nor fair to the faculty member to treat rare comments as if they are equal to ratings and comments that *are* representative of the rest of the students in a course. Evaluators need to be particularly vigilant and self-aware when they are reading or summarizing students' comments. When rare negative ratings or comments are emphasized, it presents an inaccurate picture of the students' views (Franklin & Berman, 1998; Lewis, 2001).

In many cases, it is not feasible to include *all* student comments (e.g., if the course is very large or if students provide significant written feedback). When results are summarized and only mean or median ratings are included in a dossier, negative scores and comments are inadvertently given extra weight in a review. Administrators should be careful to include comments that are representative of the students' views. Many administrators feel an obligation to include negative comments, even when they are not representative. Instead, compilers should focus on presenting a *representative* summary or sampling of students' comments. In other words, a single negative comment should not be included if it represents a miniscule proportion of the written comments and/or would misrepresent the distribution of students' comments.

One of the best ways to ensure that summaries of comments represent students' views is to sort student comments into groups based on similarity and label the group with a theme (Lewis, 1991), then rank the themes based on the frequency of comments in each (see Figure 1). Note that many students include multiple topics in a single sentence so those should be broken into topical fragments and each sorted separately. Faculty members should focus improvement efforts on the first two to three themes, not the most negative comment(s). Some common themes include Labs, Homework, Teamwork, Lecture, Availability, Textbook, and Exams. Sorting written comments by theme not only helps highlight which comments are frequent and rare, it helps reviewers and faculty not to overemphasize isolated comments, whether positive or negative.

Figure 1. Sample format for a thematic analysis of students' written comments. Available at http://www.schreyerinstitute.psu.edu/tools/?q=template.



Faculty Name **Course Name Semester Year**
Section # **Course Number**# responses / # enrolled

Summary of Student Comments

This is a template for analysis of student comments.
The themes below are examples of themes—they are not inclusive. Different themes will emerge for each course and instructor.

1. ***What helped you learn in this course?***

Class Discussion
-
-
-
-
-

Instructor Knowledge
-
-
-
-

Instructor Style / Enthusiasm / Approachability
-
-
-

Groupwork/Teamwork
-
-

Teaching Methods
-

Homework
-
-

Readings
-

Misc.

Faculty Name **Course Name Semester Year**
Section # **Course Number**# responses / # enrolled

Identify common themes within students' answers to each question.
Order the themes from those with the most comments to those with the least to identify the most critical issues for the most students.

2. ***What suggestions do you have for changes that would improve your learning?***

Organization
-
-
-
-
-

Workload
-
-
-
-
-

Clarify Expectations
-
-
-
-

Assignments
-
-
-

Grading/Grading Criteria
-
-
-

Lectures

That said, the student ratings research community has repeatedly voiced concerns about students' written comments being included in personnel decisions because they duplicate the information from the same students who have completed the ratings (Franklin & Berman, 1998). Arreola (2007) considers students' written comments to be subjective and unreliable. Marsh (2007) provides an overview of the research on written comments, which is relatively small, but does indicate alignment between written comments and student ratings.

8. Contradictory written comments are not unusual.

It is a rare faculty member who does not receive at least some contradictory comments in the written feedback that typically accompanies student ratings (Marincovich, 1999). Neither administrators nor review committee members should consider this diagnostic. Administrators typically recognize that the situation is common because they see many more student ratings reports than do faculty who serve on review committees. New faculty can be particularly frustrated or concerned when students' comments contradict each other given that they generally feel additional pressure to perform well on student ratings because they feel that their tenure decision or their reappointment depends on uniformly good student ratings and comments. Administrators and faculty who have served on review committees can help their junior peers focus on the most frequent ratings and comments.

Questions Asked by Administrators about Student Ratings: Providing Feedback and Responding to Faculty Concerns

Administrators, and sometimes faculty review committees, are responsible for providing useful and actionable feedback to guide faculty career development in pre-tenure reviews or reappointments. Both administrators and reviewers can experience discomfort with making life-altering decisions about other faculty based on student ratings data (though hopefully not solely on those data). The discomfort can be exacerbated if these individuals do not know about the history of student ratings at the institution, if they are unfamiliar with the research literature, or if they have been operating under misconceptions. Below are some of the most common questions asked by faculty and administrators; these are not just questions asked by faculty who receive low ratings or are unhappy with their results.

1. *How do I know whether a faculty member's ratings are "good" or "bad"?*

Look at the distribution of the ratings across the scale, not solely at the mean or the median. Most student ratings distributions are skewed, i.e., not normally distributed, with the peak of the distribution above the midpoint of the scale. The mean misrepresents the ratings in a skewed distribution because a few low ratings in the tail of the distribution can pull the mean down. It is unacceptable to allow a faculty member "to be portrayed as a less effective teacher with lower ratings" (Berk, 2013, p. 74) because of an institution's choice of which measurement of central tendency to report. Distributions that include the ratings of multiple faculty for the purposes of improving the teaching or curriculum within a department, degree program, or course can provide useful comparative information (Arreola, 2007; Berk, 2013; Hativa, 2013).

Most institutions in the U.S. use a norm-referenced approach to interpreting a faculty member's ratings (Hativa, 2013b; McKeachie, 1997). For example, faculty with most of their ratings distributed across scores of 3.5–5 on a 5-point scale (or 5–7 on a 7-point scale) are doing well, even if they have a few stray scores in the lower ratings. If a large percentage of the ratings are clustered at the higher end of the scale, the faculty member is doing fine—even if a few students rate the faculty member at the low end of the scale. Student ratings are intended to represent the collective views of students, not the rare views. Even when a faculty member is doing fine, her/his history of ratings may include a couple of courses that were rated lower. Every faculty member receives some lower ratings at some point in her/his career.

Faculty members with a normal distribution of scores or a distribution with the peak below the midpoint of the scale likely have an instructional issue (or issues) that need attention (Arreola, 2007). The issues may be easily addressed or may be more serious, but all faculty members should be given the opportunity to address students' concerns. In other words, do not ignore low scores!

2. *What should I say to a faculty member with ratings distributed across the low end of the rating scale?*

Faculty with many scores in the 1–2 range on a 5-point scale (or 1–3 range on a 7-point scale) or with scores relatively evenly distributed across the entire scale are typically facing serious challenges with their students. These kinds of distributions need to be addressed as soon as possible. Faculty members who receive these kinds of rating distributions in most of their courses need sufficient time to develop their teaching before coming up for a formal evaluation or a contract renewal.

These faculty members should also be reassured that even though some faculty seem "born to teach," nearly all of the behaviors practiced by excellent teachers can be learned. Faculty members with low ratings should be reminded of the ways that the college or university provides support for effective teaching as well as online and library resources on effective teaching in higher education. Recommend that the faculty work with a senior faculty member who is a good teacher and mentor, or remind her/him of other resources that excellent faculty use, such as the resources provided by the campus teaching center (Wilson, 1986). The senior faculty member must be a good mentor as well as a good teacher because good mentors do not simply expect a mentee to copy her/his teaching.

If a pattern of low scores develops, the faculty member should be encouraged to seek mentoring, coaching, or advice from a professional in the campus teaching and learning center. Research indicates that faculty who work with an expert or knowledgeable colleague (one who does not simply say, "Teach like me") do improve (Boice, 2001; Geis, 1991; Brinko, 1991). However, faculty should not simply be "sent to the teaching center" in response to low or problematic student ratings because the teaching center should not be seen as a punishment, but as a support offered by the university. It is far better to begin talking with faculty immediately upon their arrival on campus about the resources the institution provides as a way to ensure that all faculty are successful teachers.

Most teaching centers practice confidentiality with their faculty clients (cf. http://podnetwork.org/about-us/pod-governance/ethical-guidelines/). This means that even if an administrator recommends that a faculty member seek help from the teaching center, center personnel will not report back to the administrator about that consultation (Zakrajsek, 2010). Administrators are free to refer faculty to contact the teaching center, but most centers will treat the faculty member as if she/he voluntarily sought consultation. Administrators generally respond positively to these traditions and are more concerned that their faculty members be treated with respect and dignity than they are about getting a report from the center. Rather than request a follow-up from the center, administrators can take a more constructive approach by asking to meet with the faculty member at a future point to discuss what she/he has been doing to improve their teaching and address students' concerns. Many centers also provide consultation services to administrators who are seeking advice about how to mentor faculty within their units.

3. *How do I respond to a faculty member who says that "only faculty who give away 'A' grades get high ratings" or who argues that another faculty member who receives high ratings "must be giving away grades"?*

Most faculty members at most institutions receive high student ratings (Arreola, 2007; Hativa, 2013a). Every institution has numerous examples of faculty with high academic standards who also receive high student ratings. Administrators may want to share the departmental or course distribution (as opposed to simply the departmental average) as a way for faculty members to calibrate their own results. Some faculty respond better to a conversation with a respected faculty member in the department who is tough, but fair, and who also receives high ratings; most departments have at least one such faculty member.

Student ratings researchers have long been studying the relationship between grades and ratings (Abrami et al., 1980; Eiszler, 2002; Marsh, 1987). While a number of studies have shown no relationship between grades (or expected grades) and student ratings (Marsh & Roche, 1997; Gigliotti & Buchtel, 1990), more research studies document that students' grades are positively correlated with student evaluations (Abrami, 2001; Eiszler, 2002; Feldman, 1976). The most commonly cited correlation is 0.2–0.3, but researchers report correlation coefficients that vary from 0.1–0.5 (Abrami et al., 1980; Arreola, 2007; Feldman, 1976; Greenwald & Gilmore, 1997; Stumpf & Freedman, 1979). Marsh (2007) suggests that the majority of the research indicates support for the hypothesis that students who learn more earn higher grades and give higher ratings. More recently, Benton and colleagues have documented that students give instructors higher ratings when students are expected to take on some share of responsibility for learning (Benton, Guo, Li, & Gross, 2013 as cited in Benton & Li, 2013).

The positive though weak correlation leads researchers to recommend that evaluators use extreme caution when inferring that a faculty member's grading policy has significantly impacted their ratings. The combination of high ratings and higher grades might represent student learning, grading leniency, or students' characteristics unrelated to instruction (McKeachie, 1979, 1997). None of the stories that claim grading practices are responsible for grade inflation is widely accepted by the student ratings research community. In fact, McKeachie (1990) notes that faculty members who are effective working with poorer students receive higher ratings from those students than they receive from other students.

Most students do not equate faculty who have high standards with poor teaching. Faculty members who try to manipulate students' ratings by "giving away As" should be advised that they are at risk of receiving low ratings from students who worked hard in the course and who turned in A work (Abrami et al., 1980; McKeachie, 1997). In other words, poor teachers who try to increase their scores by boosting grades are unlikely to fool students.

In a similar vein, some faculty members suggest that their low ratings are a result of "high standards" and students' dislike of homework or even a reasonable workload. A heavy workload is not always synonymous with "academic rigor" (Franklin, 2001), so an over-ambitious workload could reasonably result in lower student ratings. Peer review of faculty teaching materials such as syllabi and assignments, course observations (Chism, 2007), and review of students' work (Cashin, 1995) are the best methods for evaluators to determine

whether a faculty member is expecting too much or too little from students and whether students are earning undeserved high grades.

4. *How do I respond to a faculty member who says that student ratings are "just a popularity contest" and that they are "not valid"?*

As noted above, the purpose of student ratings is to gather students' perspectives on the instruction or learning environment in a course (Hativa, 2013a). Their validity has been tested more than any other method for evaluating faculty teaching (Abrami, 2001; Abrami et al., 1990; Aleamoni, 1999; d'Appolonia & Abrami, 1997; Feldman, 1989; Marsh, 1984, 1982b; Marsh & Roche, 1997). The majority of the legitimate research on student ratings indicates that they are a more reliable and valid representation of teaching quality than any other method of evaluating teaching, including peer observation, focus groups, and external review of materials (McKeachie, 1997; Berk, 2005, 2013) and they are highly correlated with other measures of teaching effectiveness (Abrami et al., 1990; Berk, 2013). In other words, most other methods for evaluating teaching have not undergone extensive statistical analyses or been exposed to the scrutiny of student ratings, yet we continue to rely on them.

When faculty question the *validity* of students' ratings, they are typically not concerned about the statistical validity of the ratings instrument of the results. Instead, faculty who question the validity of student ratings are generally most concerned that students view are wrong, but that their ratings being used against the faculty member. This provides an opportunity to talk about the value of students' views and that student ratings are just one source of data in the faculty evaluation process.

Be honest that student ratings are unlikely to become obsolete any time soon, no matter what the latest headlines say. Because student ratings provide an effective and systematic way to gather feedback from students enrolled in courses, it is in the faculty member's best interest to learn how to use these data to benefit his/her. Specifically, instructors who want to increase their ratings should focus their efforts on improving the learning environment for students through by building "communication, motivational, and rapport-building skills" (IDEA Research Note 1, 2003). Campus teaching and learning centers have many resources and strategies to help faculty develop these and other attributes of effective teaching. At Penn State, faculty from every campus and college may seek the services of Schreyer Institute for Teaching Excellence. To find out which of the Institute's faculty consultants serve your academic unit, visit http://www.schreyerinstitute.psu.edu/Help/Liaisons/.

5. *What should I say when a faculty member argues that students are biased against him/her?*

Students, like all human beings are biased. But students, like other members of society, are not monolithic in their views. In other words, not all students are biased in the same ways. The real question here is whether student bias against some attribute of a faculty member is widespread and strong enough to overwhelm the students' ratings of the faculty member's teaching or course environment and solely reflect students' bias.

Faculty who do not fit students' perceptions of what a professor should look or act like can experience bias from the students. Student ratings researchers have identified among students the same biases that exist in society (gender, sexual orientation, political, religious, etc.). While these biases definitely exist, the research indicates that the biases rarely, if ever, fully explain ratings that cluster at the low end of the ratings scale.

The fact that student ratings instruments are not designed to capture rare student views is one reason why we hear contradictory information about whether or not student ratings are biased against women faculty, faculty of color, and other non-majority attributes of faculty. For many years, studies that analyzed large samples of courses from a variety of disciplines consistently showed no significant difference in ratings due to systematic gender bias (Feldman, 1992, 1993; Franklin & Theall, 1994). Yet, women faculty, particularly in male-dominated fields in the STEM disciplines (science, technology, engineering, and math) continued to suggest that these studies did not represent their experiences. Given the relatively small numbers of women faculty in these fields, ratings that reflect bias will be represented in the tails of the distribution, not in the peak of the distribution. As a result, these biases are more difficult to detect.

Over time, a growing body of research has documented gender effects on student ratings, but these effects are neither uniform nor consistent across all disciplines, nor do they apply to all women (e.g., Bachen et al., 1999; Basow, 1995; Centra & Gaubbatz, 2000; Hancock et al., 1993; Sinclair & Kunda, 2000). While recent stories in the academic press have generated a lot of attention, the articles cited (Bragaa et al., 2014; MacNell et al., 2015) have methodological issues, and significantly overstate the case (Ryalls et al., 2016).

The research on gender bias has a longer history than does the research on racial, ethnic, or cultural bias, in part because minority faculty still constitute a relatively small percentage of the faculty. The number of studies is increasing and evidence is mounting that such biases exist among students and may impact student ratings (Anderson & Smith, 2005; Davis, 2010; Galguera, 1998; Gilroy, 2007; Hendrix, 1998; Lazos, 2011; Reid, 2010; Smith, 2007, 2009; Smith & Hawkins, 2011; Smith & Johnson-Bailey, 2011/12). However, at this point the bias is not sufficiently strong or widespread to explain consistently low ratings across all courses for a faculty member.

6. *How should I respond to a faculty member who suggests that online administration of student ratings resulted (or will result) in lower ratings?*

Many faculty members feel that the move to online administration of student ratings has resulted in low ratings. This is generally not supported by the ratings data, i.e., ratings distributions of most faculty members continue to cluster at the high end of the scale as do most aggregate departmental and college distributions (Linse, 2010). In the early days of online student ratings, Northwestern University reported on a study (Hardy, 2003) that included both increases and decreases, as well one that showed a slight decrease (-0.25 on a 6-point scale). Faculty at The Pennsylvania State University (Penn State) had similar concerns, but one study showed only a small increase in scores of 1–3 on a 7-point scale, as well as a marked increase in

ratings of 7 (Linse, 2010; Linse & Xie, 2011). The IDEA Center,[2] which processes student ratings from hundreds of institutions, reports no difference in online ratings (Webster et al., 2010) as do numerous other studies (Dommeyer et al., 2004; McGhee & Lowell, 2003; Stowell et al. 2012). No reports document an increase in bi-modal distributions in institutionally administered ratings. Now that online student ratings have become commonplace, it has become clear that students who are engaged in a course are more likely to complete the student ratings than students who are disengaged (Berk, 2013).

Other potential causes should be ruled out before attributing a ratings change to the method of administration, particularly because such changes are relatively rare (though not impossible). Request that the faculty member provide comparison data from paper and online student ratings distributions for the same course. If a faculty member has not taught the course for many years, during which the transition to online happened, the results may not be directly attributable to the online transition. The course material may be out-of-date or it may rely too heavily on out-of-date teaching methods. Students today expect at least some level of engagement in class, in both face-to-face and online courses (Barkley, 2010).

Some individual faculty members may be able to make a case that their ratings changed dramatically before and after the shift to online administration. When this can be substantiated, the department or program head should include a note in the faculty member's dossier.

7. *How do I tell a long-serving faculty member who has had poor student ratings for years that those ratings are no longer acceptable?*

Poor student ratings may have been acceptable in the past, but the issue may also have been avoided for other reasons including not knowing what kind of ratings are acceptable, not knowing how to approach the faculty, or wanting to avoid hurting or discouraging the faculty member (Gunsalus, 2006).

The administrator can ease into the conversation by saying, "It may have been sufficient in the past to receive these kinds of ratings, but things have changed and students expect more now. The university has invested resources to help you take the next steps to improve your teaching. For example, …" Most colleges and universities have a variety of resources to support faculty professional development including experienced teaching mentors, faculty learning communities (Cox, 2004), and teaching and learning centers (Brinko, 1991; Ouellett, 2010; Sorcinelli & Austin, 2006; Sorcinelli et al., 2006).

8. *How do I respond to faculty who have been told that "teaching doesn't matter for promotion and tenure (P&T)"?*

At many colleges and universities, it is true that faculty cannot expect to be successful in the promotion and tenure process based on excellent teaching and mediocre research (Glassick et

2. IDEA used to be an acronym for "Instructional Development & Evaluation Assessment," a student ratings form developed at Kansas State University. The phrase behind the acronym is no longer used by the IDEA Center and does not appear on their website (http://www.ideaedu.org/) as of November 19, 2016. In other words, IDEA is no longer an acronym.

al.1997; Fairweather, 2002; Soderberg, 1985). In the U.S., faculty on the tenure track at nearly all institutions (except tenure-line faculty at community colleges), have research responsibilities in addition to teaching and service responsibilities. At research-focused universities in particular, a largely unwritten rule exists that unless faculty research productivity is acceptable, they will not seriously be considered for tenure. Miller and Seldin (2014, p. 1) note that the importance of research and publication continues to increase in the faculty evaluation process, which appears to support the "observation that faculty members are paid to teach but are rewarded for their research and publication."

There was once great hope that the Scholarship of Teaching and Learning (SOTL; Boyer, 1990) would evolve so that scholarly teaching would "count" for more in the promotion and tenure process (Huber, 2002). Things have changed at some institutions so that SOTL does "count" in promotion and tenure decisions, but primarily when the SOTL has been published in peer-reviewed journals and/or resulted in grant support.

Today, what has changed is that poor teaching can now have a significant negative impact on a tenure and/or promotion case. This is particularly true if the faculty member does not have a strong research record, whether disciplinary or SOTL. This change is, in part, a result of Boyer's and others' work to broaden the definition of scholarship, but also because of tightening budgets, higher tuition, and increased calls for accountability. The bottom line is that in today's world, few faculty members can afford to ignore teaching, not even "star researchers."

9. *What do I say to a faculty member who says, "My response rates are too low to be included in my dossier"?*

Unless an institution has a set minimum response rate for inclusion in the dossier, all results will need to be included. There is no single standardized "ideal" response rate although a number of researchers have made suggestions (Franklin & Theall, 1991; Marsh, 1984; Nulty, 2008; the recommendations of the latter are reproduced by Barre, 2015). These recommended response rates are challenging to obtain for online student ratings, but it takes greater effort on the part of faculty to achieve high response rates. Response rates for online administration tend to fall by 25–30% below those of paper student ratings (Benton et al., 2010; Hativa, 2013a; Johnson, 2003; Nulty, 2008; Sorensen & Reiner, 2003). Response rates may rebound as students no longer expect paper student ratings and mobile versions allow in class administration.

Ultimately, faculty members will need to trust that their colleagues will be skeptical that results from extremely low-response courses are representative of students' views. That said, colleagues and administrators are unlikely to tolerate extremely low response rates over multiple years, particularly since all faculty can implement at least some of the strategies known to boost response rates (Berk, 2006; Nulty, 2008). Effective strategies include discussing the importance of student ratings to the faculty member and describing faculty efforts to use student feedback to improve the course, noting that student feedback will benefit future students, and multiple reminders from the faculty. Many online systems are programed to provide automatic reminders when a student has unrated courses; at Penn State students receive reminders only if they have unfinished SRTEs. Some faculty have had great success in rewarding students for reaching a particular response rate or providing extra credit points (Dommeyer et al., 2004), but other

faculty feel strongly that such rewards amount to bribery for higher ratings. Two extremely successful practices are 1) granting students early access to grades or 2) granting access to results; the former may not be technologically possible and some faculty feel strongly that students should not see the results, especially when those results are used in personnel decisions. See http://www.schreyerinstitute.psu.edu/IncreaseSRTERespRate/ for the results of an informal study in which faculty described what they do to receive response rates at or above 70%.

A number of efforts can help, including repeated reminders from the online system, reminders from faculty, and sincere comments from faculty that their responses will be read and taken seriously (Nulty, 2008). Faculty members may also want to consider regularly collecting feedback from students during the term, which creates a habit of feedback and builds trust among students that the faculty member is sincere in his/her respect for students' perspectives (Svinicki, 2001).

Some institutions have policies that allow faculty who want to experiment with new teaching methods or new course content to arrange in advance to exclude the student ratings for the experimental course from the faculty member's dossier. For example, Penn State's Statement of Practices for the Evaluation of Teaching Effectiveness for Promotion and Tenure states, "If there is some reason to explain the results or the absence of results in a particular case, the appropriate academic administrator shall make a note to that effect in the dossier. For example, in advance of a course being taught for the first time in an experimental way, an administrator and a faculty member might agree not to administer the SRTE [Student Ratings of Teaching Effectiveness]. Such agreements should be in writing" (http://vpaa.psu.edu/files/2016/09/srte_statement-248pj9j.pdf). Other universities have similar language in their reappointment, promotion, and tenure (RPT) policies. We suggest that the student ratings be administered even if an administrator agrees to the exclusion because some faculty have found that their ratings do not decrease as expected.

10. How do I respond to faculty who say that the lower response rates of the online student ratings system make the ratings "invalid"?

As noted above, the validity of student ratings has been well established for decades. When some faculty express concerns about validity, they are actually concerned about the representativeness of the sample of responding students, not the statistical validity of the instrument. Faculty are wise to be concerned about the response rate as smaller numbers of responses are less likely to be representative (Benton et al., 2010; Berk, 2013). As noted above, average response rates typically decrease with the transition to online ratings. However, no research has reported a systematic or widespread decrease in average or median ratings and some have reported stable or increased averages (Ardalan et al., 2007; Dommeyer et al., 2004; Hardy, 2003; Venette et al., 2010)

Some institutions have begun to see response rates rebound as students become more accustomed to online ratings and as students who have experienced paper administration graduate (Johnson, 2003). Other institutions have been able to increase response rates by offering student respondents access to the results, early access to grades, or mobile versions of the online system (Berk, 2012; Kaplan, 2014). Many faculty have found success emphasizing how

important the feedback is to the improvement of the course and by providing examples of course improvements suggested by past students; for some of these strategies, see http://www.schreyerinstitute.psu.edu/IncreaseSRTERespRate/.

Faculty with low response rates in small-enrollment courses may have cause for concern because when the number of respondents is small, a single student's rating carries a lot of weight. But as noted above, the lower response rates have typically not had a negative impact on faculty members' average scores. Administrators should be wary of over-interpreting small-enrollment courses with low response rates.

Conclusion

The conclusions of research experts in the field of student ratings are not reaching the faculty and administrators who are responsible for faculty evaluation. Too often, faculty misperceptions about student ratings are instead obtained from the academic, and sometimes mainstream, press which largely ignores the more than 80 years of research on the topic. Second, student ratings are so important in the faculty evaluation process, especially in terms of personnel decisions, that we can no longer afford to ignore the misuse and misinterpretation of student ratings data.

While the two final sections of this report are written for different audiences, both focus on one important issue—that the appropriate use of student ratings data is fundamental to building a high-quality teaching ecosystem within an institution. Inappropriate use of student ratings breeds mistrust, fosters inequities and inconsistencies, and ultimately demoralizes the faculty. With increased appropriate and accurate use of student ratings data, faculty and administrators can begin to avoid other unintended consequences such as turning the important process of listening to students' voices into a rote activity that has no meaning for the students or the faculty.

Research-based decisions can help to create a more coherent academic community that is empowered to take responsibility for high-impact work on campus. If student ratings data are used appropriately, faculty once closed to or dismissive of students' feedback may be able to approach student ratings from a more open-minded perspective. A greater understanding of student ratings could lead to broader appreciation within the faculty community of faculty whose primary responsibility within the community is to help the institution meet its mission of educating students.

References

Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. *New Directions for Institutional Research*, *109*, 59–87.

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990).The validity of student ratings of instruction: What we know and what we don't. *Journal of Educational Psychology*, *82*(2), 219–231.

Abrami, P. C., Dickens, W. J., Perry, R. P., & Leventhal, L. (1980). Do teacher standards for assigning grades affect student evaluations of instruction? J*ournal of Educational Psychology*, *72*,107–118.

Aleamoni, L. M. (1999). Student rating myths versus research facts: An update. *Journal of Personnel Evaluation in Education*, *13*(2), 153–166.

Ardalan, A., Ardalan, R. Coppage, S., & Crouch, W. (2007). A comparison of student feedback obtained through paper-based and web-based surveys of faculty teaching. *British Journal of Educational Technology*, *38*(6), 1085–1101.

Anderson, K. J., & Smith, G. (2005) Students preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Sciences*, *27*(2), 184–201.

Arreola, R. A. (2007). *Developing a comprehensive faculty evaluation system*, 3rd ed. Bolton, Massachusetts: Anker.

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999) Assessing the role of gender in college students' evaluations of faculty. *Communication Education*, *48*(3, July), 193–210.

Barkley, E. F. (2010). Student engagement techniques: A handbook for college faculty. San Francisco, California: Jossey-Bass.

Barre, B. (2015). Academic blogging and student evaluation click bait: A follow-up. Reflections on Teaching and Learning, the CTE Blog. Center for Teaching Excellence, Rice University. http://cte.rice.edu/blogarchive/2015/07/28/studentevaluationsfollowup .

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology*, *87*, 656–665.

Benton, S. L. & Cashin, W. E. (2011). Student ratings of teaching: A summary of research and literature. IDEA Paper No. 50. Center for Faculty Education and Development. IDEA Center, Kansas State University. http://ideaedu.org/wp-content/uploads/2014/11/ideapaper_50.pdf .

Benton, S. L., Guo, M., Li, D., & Gross, A. (2013, April). Student ratings, teacher standards, and critical thinking skills. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, California.

Benton, S. L. & Li, D. (2015) Response to "A Better Way to Evaluate Undergraduate Teaching" IDEA Editorial Note #1, IDEA Center. http://ideaedu.org/research-and-papers/editorialnotes/response-to-wieman/.

Benton, S. L., Webster, R., Gross, A. B., & Pallett, W. H. (2010). An analysis of IDEA student ratings of instruction using paper versus online survey methods, 2002-2008 data. IDEA Technical Report, No. 16. http://ideaedu.org/wp-content/uploads/2014/11/techreport-16.pdf.

Berk, R. A. (2013). *Top 10 flashpoints in student ratings and the evaluation of teaching: What faculty administrators must know to protect themselves in employment decisions*. Sterling, Virginia: Stylus.

Berk, R. A. (2012). Top 20 strategies to increase the online response rates of student rating scales. *International Journal of Technology in Teaching and Learning*, *8*(2), 98–107.

Berk, R. A. (2006). *Thirteen strategies to measure college teaching: A consumer's guide to rating scale construction, assessment, and decision making for faculty, administrators, and clinicians*. Sterling, Virginia: Stylus.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*(1), 48–62.

Boice, R. (2001). *Advice for new faculty members: Nihil Nimus*. Boston, Massachusetts: Allyn and Bacon.

Boyer, E. L. (1990). *Scholarship reconsidered: Priorities of the professoriate*. Princeton, New Jersey: Carnegie Foundation for the Advancement of Teaching.

Bragaa, M., Paccagnellab, M., & Pellizzaric, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review 41*, 71–88.

Braskamp, L. A., Brandenburg, D. C., & Ory, J. C. (1984). Evaluating teaching effectiveness: A practical guide. Beverly Hills, California: Sage.

Braskamp, L. A., Ory, J. C., & Pieper, D. M. (1981). Student written comments: Dimensions of instructional quality. *Journal of Educational Psychology*, *73*(1), 65–70.

Brinko, K. T. (1991). The interactions of teaching improvement. *New Directions for Teaching and Learning*, 48, 21–37.

Cashin, W. E. (2003). Evaluating college and university teaching: reflections of a practitioner. In, *Higher education: handbook of theory and research*, edited by J. C. Smart, pp. 531–593. Dordrecht, The Netherlands: Kluwer Academic.

Cashin, W. E. (1999). Student ratings of teaching: uses and misuses. In, *Changing Practices in Evaluating Teaching: A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions*, edited by P. Seldin, pp. 25-44. Bolton, Massachusetts: Anker.

Cashin, W. E. (1996). Developing an Effective Faculty Evaluation System. IDEA Paper No. 33. http://ideaedu.org/wp-content/uploads/2014/11/Idea_Paper_33.pdf.

Cashin, W. E. (1995). Student ratings of teaching: The research revisited, IDEA Paper No. 32. http://files.eric.ed.gov/fulltext/ED402338.pdf.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education*, *71*(1), 17–33, January-February 2000.

Chism, N V. (2007). *Peer review of teaching: A sourcebook*. Bolton Massachusetts: Anker.

Cox, M. D. (2004). Introduction to faculty learning communities. New Directions for Teaching and Learning, *97*, 5–23.

d'Appolonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist*, *52*(11), 1198-1208.

Davis, D. J. (2010). The experiences of marginalized academics and understanding the majority: Implications for institutional policy and practice. *International Journal of Learning* 17(6), 355– 364.

Dommeyer, C. J., Baum, P., Hanna, R. W. & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: Their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education*, *29*(5), 611–623.

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education*, *43*(4), 483–501.

Fairweather, J. S. (2002). The ultimate faculty evaluation: Promotion and tenure decisions. *New Directions for Institutional Research, 114*, 97–108.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: evidence from student ratings, reprinted in *The scholarship of teaching and learning in higher education: An evidence-based perspective*, edited by R. P. Perry & J. C. Smart, pp. 93–95.

Feldman, K. A. (1993). College students' views of male and female college teachers, Part II--Evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, *34*(2), 151–211.

Feldman, K. A. (1992). College students' views of male and female college teachers: Part I—Evidence from the social laboratory and experiments. *Research in Higher Education, 33*(3), 317–375.

Feldman, K. A. (1989). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education, 30*(2), 137–189.

Feldman, K. A. (1976) Grades and college students' evaluations of their courses and teachers. *Research in Higher Education, 4*, 69–111.

Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. *New Directions for Teaching and Learning, 87*, 85–100.

Franklin, J., & Berman, E. (1998). Using student written comments in evaluating teaching. *Instructional Evaluation and Faculty Development, 18* (1) [text version in possession of the author has no page numbers].

Franklin, J., & Theall, M. (1994). "Student ratings of instruction and sex differences revisited." Paper presented at the 78th annual meeting of the American Educational Research Association, New Orleans. April 7, 1994.

Franklin, J. L., & Theall, M. (1991). Grade inflation and student ratings: a closer look. Paper presented at the 72nd annual meeting of the American Educational Research Association. Chicago: April 7. ERIC # ED 349 318.

Galguera, T. (1998). Students' attitudes towards teachers' ethnicity, bilinguality, and gender. *Hispanic Journal of Behavioral Sciences*, *20*(4), 411–429.

Geis, G. L. (1991). The moment of truth: Feeding back information about teaching. *New Directions for Teaching and Learning*, *48*, 7–19.

Gigliotti, R. J., & Buchtel, F. S. (1990). Attributional bias and course evaluations. *Journal of Educational Psychology*, *82*(2), 341–351.

Gilroy, M. (2007) Bias in Student Evaluations of Faculty? *The Hispanic Outlook in Higher Education 17*(19), July 2, 2007, 26–27.

Glassick, C. E., Huber, M. T., & Maeroff, G. I. (1997). *Scholarship assessed: Evaluation of the Professorate*. San Francisco, California: Jossey Bass.

Greenwald, A. G., & Gillmore, G. M. (1997) Grading lenience is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209–1217.

Gunsalus, C.K. (2006). The college administrator's survival guide. Cambridge, Massachusetts: Harvard University Press.

Gutgold, N. D., & Linse, A. R. *Women in the academy: Learning from our diverse career pathways*. Lanham, Maryland: Lexington.

Hancock, G. R., Shannon, D. M., & Trentham, L. L. (1993). Student and teacher gender in ratings of university faculty: results from five colleges of study. *Journal of Personnel Evaluation in Education, 6*(3), 235–248.

Hardy, N. (2003). Online ratings: fact and fiction. *New Directions for Teaching and Learning, 96*, 31–38.

Hativa, N (2013a). *Student ratings of instruction: A practical approach to designing, operating, and reporting*. Oron Publications.

Hativa, N. (2013b). *Student ratings of instruction: recognizing effective teaching*. Oron Publications.

Hendrix, K. J. (1998) Student perception of the influence of race on professor credibility. *Journal of Black Studies, 28*(6), 738–763.

Huber, M. T. (2002). Faculty evaluation and the development of academic careers. New *Directions for Institutional Research*, 114, 73–83.

Husbands, C. T. (1997). Variations in students' evaluations of teachers' lecturing in different courses on which they lecture: A study at the London School of Economics and Political Science. *Higher Education 33*, 51–70.

IDEA Research Note 1. (2003). The "excellent teacher" item. Manhattan, Kansas: The IDEA Center.

Johnson, T. D. (2003). Online student ratings: will students respond? *New Directions for Teaching and Learning*, *96*, 49–59.

Kaplan, M. (2014). *Release of Course Evaluations to Students, Policies of University of Michigan Peer Institutions.* Center for Research on Learning and Teaching, University of Michigan. October 2014.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research, 109,* 9–25.

Lazos, S. R. (2011) Are student teaching evaluations holding back women and minorities? The perils of "doing" gender and race in the classroom. In, *Presumed Incompetent: The Intersections of Race and Class for Women in Academia*, edited by G. y M. Gabriella et al., Chapter 12. Boulder, Colorado: Utah State University Press, an imprint of Colorado University Press.

Lewis, K. G. (2001). Making sense of student written comments. *New Directions for Teaching and Learning*, *87*, 25–32.

Lewis, K. G. (1991). Gathering data for the improvement of teaching: What do I need and how do I get it? *New Directions for Teaching and Learning*, *48*, 65–82.

Linse, A. R. (in press). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation* (2017); http://dx.doi.org/10.1016/j.stueduc.2016.12.004.

Linse, A. R. (2010). Analysis of online SRTE data from select semesters (2009-2010). Prepared for the Committee on Faculty Affairs of the University Faculty Senate. Schreyer Institute for Teaching Excellence, The Pennsylvania State University, Fall 2010.

Linse, A. R., & Xie, H. (2011). Student ratings of teaching effectiveness: Analysis of data from common courses from select semesters (2009-2010). Schreyer Institute for Teaching Excellence, The Pennsylvania State University, Spring 2011.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, *40*, 291–303.

Marincovich, M. (1999). Using student feedback to improve teaching. In, *Changing practices in evaluating teaching: a practical guide to improved faculty performance and promotion/tenure decisions*, edited by P. Seldin, pp. 45–69. Bolton, Massachusetts: Anker.

Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In, *The scholarship of teaching and learning in higher education: An evidence-based perspective*, edited by P. Perry & J. C. Smart, pp.319–384. New York, New York: Springer.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research 11*, 253369.

Marsh, H. W. (1984). Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology 76*(5), 707–754.

Marsh, H. W. (1982a). Factors affecting students' evaluations of the same course taught by the same instructor on different occasions. *American Educational Research Journal*, *19*(4, Winter), 485– 497.

Marsh, H. W. (1982b). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology*, 1982, *74*, 264–279.

Marsh, H. W. (1980). Research on students' evaluations of teaching effectiveness. *Instructional Evaluation*, *4*(5), 5–13.

Marsh, H. W. & Dunkin, M. J. (1992). Students' evaluations of university teaching: A munltidimentional perspective. In, *Higher education: handbook of theory and research*, Volume 8, edited by C. J. Smart, pp. 143–233. New York, New York; Agathon.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist*, *52*, 1187–1197.

McGhee, D. E., & Lowell, N. (2003). Psychometric properties of student ratings of instruction in online and on-campus courses. *New Directions for Teaching and Learning*, *96*, 39–48.

McKeachie, W. J. (1997). Student ratings: the validity of use. *American Psychologist*, *52* (11), 1218– 1225.

McKeachie, W. J. (1990). Research on college teaching: the historical background, *Journal of Educational Psychology 82*(2), 189–200.

McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe 65*(6), 384–397.

Miller, J. E. & Seldin, P. (2014) Changing practices in faculty evaluation: Can better evaluation make a difference? *Academe 100*(3), 35–38, May/Jun 2014.

National Academies (2006) *Beyond bias and barriers: Fulfilling the potential of women in academic science and engineering*. Committee on maximizing the potential of women in academic science and engineering and Committee on science, engineering and public policy. Washington, DC: National Academies.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment and Evaluation in Higher Education 33*(3, June), 301–314.

Ory, J. C. (2001) Faculty thoughts and concerns about student ratings. *New Directions for Teaching and Learning*, *87*, 3–15.

Ory, J. C., Braskamp, L. A., & Pieper, D. M. (1980). Congruency of student evaluative information collected by three methods. *Journal of Educational Psychology*, *72*, 181–185.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? *New Directions for Institutional Research, 109*, 27–44.

Ouellett, M. L. (2010). Overview of faculty development: History and choices. In, *A guide to faculty development*, 2nd ed., edited by K. Gillespie and D. Robertson, pp. 3–20. San Francisco, California: Jossey-Bass.

Pallett, W. H. (2006). Uses and abuses of student ratings. In, *Evaluating faculty performance*, edited by P. Seldin, pp. 50–65). Bolton, Massachusetts: Anker.

Remmers, H. H. (1933). Learning, effort, and attitudes as affected by three methods of instruction in elementary psychology. *Purdue University Studies in Higher Education* (Monograph No. 21).

Remmers, H. H., & Brandenburg, G. C. (1927) Experimental Data on the Purdue Rating Scale for Instruction. *Educational Administration and Supervision*, *13*, 519–527.

Reid, L. D. (2010) The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education, 3*(3), 137–152.

Ryalls, K., Benton, S., Barr, J., & Li, D. (2016) Response to "bias against female instructors," IDEA Research and Papers, Editorial Notes.

Seldin, P. (1999). Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions. Bolton, Massachusetts: Anker.

Sinclair, L., & Kunda, Z. (2000). Motivated stereotyping of women: She's fine if she praised me but incompetent if she criticized me, *Personality and Social Psychology Bulle*tin, *26*(11), 1329–1342.

Smith, B. P., (2009), Student ratings of teaching effectiveness for faculty groups based on race and gender. *Education 129*(4), 615–624.

Smith B. P. (2007) Student ratings of teaching effectiveness: An analysis of end-of-course faculty evaluations. *College Student Journal*, *41*(4), 788–800.

Smith, B. P., & Hawkins, B. (2011) Examining student evaluations of black college faculty: Does race matter? *The Journal of Negro Education*, *80*(2), 149–162, Spring 2011.

Smith, B. P., & Johnson-Bailey, J. (2011/2012). Implications for non-white women in the academy. *The Negro Educational Review*, *62* & *63* (1–4), 115–140.

Soderberg, L. O. (1985). Dominance of research and publication: An unrelenting tyranny. *College Teaching*, *33*, 169–172.

Sorcinelli, M. D., & Austin, A. E. (2006). Developing faculty for new roles and changing expectations. Effective Practices for Academic Leaders, *1*(11), 1–16.

Sorcinelli, M. D., Austin, A. E., Eddy, P. L., & Beach, A. L. (2006). *Creating the future of faculty development: Learning from the past, understanding the present*. Bolton Massachusetts: Anker.

Sorenson, D. L, & Reiner, C. (2003). Charting the uncharted seas of online student ratings of instruction. *New Directions for Teaching and Learning*, *96*, 1–24.

Stowell, J. R., Addison, W. E., & Smith, J. L. (2012). Comparison of online and classroom-based student evaluations of instruction. *Assessment and Evaluation in Higher Education*, *37*(4), 465–473.

Street, S., Kimmel, E., & Kromrey, J.D. (1996). Gender role preferences and perceptions of university students, faculty, and administrators. *Research in Higher Education 37*(5), 615–632.

Stumpf, S. A., & Freedman, R. D. (1979) Expected grade covariation with student ratings of instruction: Individual versus class effects. *Journal of Educational Psychology*, *71*, 293–302.

Svinicki, M. D. (2001). Encouraging your students to give feedback. *New Directions for Teaching and Learning*, *87*, 17–24. San Francisco, California: Jossey-Bass.

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research*, *109*, 45–56.

Theall, M., & Franklin, J. (2000). Creating responsive student ratings systems to improve evaluation practice. *New Directions for Teaching and Learning*, *83*, 95–107.

Theall, M., & Franklin, J. (1990). Editors Notes. *New Directions for Teaching and Learning*, *43*, 1–14.

Venette, S., Sellnow, D., & McIntyre, K. (2010). Charting new territory: Assessing the online frontier of student ratings of instruction. *Assessment & Evaluation in Higher Education*, *35*(1), 101–115.

Webster, R. J., Benton, S., & Gross, A. (2010). Online versus paper survey delivery of college student ratings of instruction. Paper presented at the 2010 American Educational Research Associate (AERA) Annual Meeting, April 30–May 4, 2010, Denver, Colorado.

Wilson, R. C. (1986). Improving faculty teaching: Effective use of student evaluations and consultants. *The Journal of Higher Education*, *57*(2), 196–211.

Zakrajsek, T. D. (2010). Important skills and knowledge. In, *A guide to faculty developmen*t, 2nd ed., edited by K. Gillespie and D. Robertson, pp. 83–98. San Francisco, California: Jossey-Bass.

Zubizarreta, J. (1999). Evaluating teaching through portfolios. In, *Changing practices in evaluating teaching*, edited by P. Seldin, pp. 162–182. Bolton, Massachusetts: Anker.

SENATE COMMITTEE ON FACULTY AFFAIRS

- Michael Bérubé, Chair
- Renee Bishop Pierce
- Blannie Bowen
- Michael Bruno
- Delia Conti
- Ann Copeland
- Peter Dendle
- Marcia DiStaso
- Christopher Giebink
- Edward Glantz
- Terrence Guay
- Betty Harper
- Sharon Holt
- Zaryab Iqbal
- Rosemary Jolly
- Lisa Kitko
- Angela Linse
- Barrie Litzky
- Michael Lobaugh
- Carolyn Mahan
- Marc McDill
- Rajen Mookerjee
- John Nousek
- David Passmore
- Mark Patzkowsky
- Geoff Scott
- Amit Sharma
- Patricia Silveyra
- Stephen Snyder
- Bonj Szczygiel, Vice-Chair
- Jane Wilburne