FINAL REPORT

# TASK FORCE ON EVALUATION OF TEACHING PERFORMANCE

Mississippi State University
Fall, 2020

## Executive Summary:

The Provost's Task Force on Evaluation of Teaching Performance offers six recommendations for ensuring that the evaluation of teaching at Mississippi State University adheres to nationally recognized best practices. The recommendations have two overarching goals. First, they are aimed at creating a fairer, more balanced, and more thorough approach to evaluating instructional effectiveness and to promoting instructional improvement. Second, they are aimed at ensuring the highest quality learning environment and classroom experience for Mississippi State University Students. The recommendations include the adoption of an institutional standard for teaching excellence, promotion of a variety of methods for evaluating instructional effectiveness, training in use of these methods for instructors and administrators, and a reconceived "Student Course Survey" that provides valuable information about students' perceptions of and experiences in their courses.

## Table of Contents:

## Task Force Membership (2020):

Dr. Tommy Anderson
Associate Dean of Undergraduate Affairs, CAS
Dr. Tracey Baham
Director, Office of Institutional Research and Effectiveness
Ms. Margaretta Campbell
Graduate Student, Officer, MSU Graduate Student Association
Dr. Tom Carskadon
Professor, Department of Psychology, CAS
Dr. Jeffrey Dean (2019 Chair)
Department Head, Biochemistry, Molecular Biology, Entomology and Plant Pathology, CALS
Dr. Andrew Mackin
Department Head, Department of Clinical Sciences, CVM
Dr. Kelly Marsh
Professor, Department of English, CAS
Faculty Fellow, Center for Teaching and Learning
Dr. Shawn Mauldin
Director, Richard C. Adkerson School of Accountancy, COB
Dr. Lyndsey Miller
Associate Professor, Interior Design Program, CAAD
Dr. Eric Moyen
Department Head, Department of Educational Leadership, COE
Dr. Ian Munn
Associate Dean, CFR
Mr. Tyler Packer
Undergraduate Student, President, MSU Student Association
Dr. Kari Reeves
Associate Dean, BCOE
Department Head, Department of Industrial and Systems Engineering
Dr. Rebecca Robichaux-Davis
Professor, Department of Curriculum, Instruction and Special Education, COE
President, Robert Holland Faculty Senate
Prof. Michael Seymour (2020 Chair)
Professor, Department of Landscape Architecture, CALS
Acting Director, Center for Teaching and Learning

## Charges:

In an email dated July 31, 2019, Dr. David Shaw (Provost) charged the Task Force on Evaluation of Teaching Performance with four tasks:

1. Develop appropriate language for revising AOP 13.15, as necessary
2. Determine whether adjustments are necessary to questions on the student evaluation of teaching effectiveness
3. Strengthen language that encourages faculty teaching assessment beyond just using student surveys
4. Develop plans to improve student surveys, increase participation with online surveys and reduce faculty concerns about fairness and bias in student responses

Immediately prior to the task force being convened, the Robert Holland Faculty Senate and MSU administration ratified an update to AOP 13.15 that added language to conform with revised IHL Policy 407.02 requiring all contents of student evaluations, including ad hoc written comments, be provided to faculty, Deans and Department Heads.

## Process:

The task force was provided a wide variety of documents on related subjects, including assessment of faculty teaching performance, best practices in the use of student surveys of teaching performance, and validity of online student survey responses, as well as current MSU student evaluation templates, other internal MSU documents and best practices documents from peer and aspirational institutions. These documents are listed under References.

With help from Dr. Shaw, the task force arranged for a March 2020 campus visit from Dr. Ginger Clark, Associate Vice Provost for Academic and Faculty Affairs and Director of the Center for Excellence in Teaching at the University of Southern California, an expert in faculty teaching assessment. Unfortunately, her visit was canceled due to travel restrictions associated with the burgeoning Covid-19 pandemic. Shutdowns and other responses to the pandemic significantly impacted subsequent task force deliberations.

For initial discussions of the assigned tasks, the task force organized into three subcommittees focusing on equitable evaluation of faculty with respect to teaching performance, issues with the current student survey instrument, and concerns about moving student surveys to an online format. An executive subcommittee was assigned to review language in AOP 13.15 based on recommendations from the other subcommittees and an ad hoc group took up the peripheral matter of new options for student reporting of faculty misbehavior in the classroom.

Following the disruption of the pandemic, the committee regrouped and worked as a whole in the fall of 2020 with a slightly altered membership due to some turnover and new assignments. The new membership reviewed a previously completed draft of the report and recommendations and met a number of times to discuss various aspects of the issues via Webex. The recommendations and AOP 13.15 proposed revision were revised based upon these discussions and finally voted upon and approved by the task force. This final report provides the background and other information to support these

recommendations and also explains the major findings and a related discussion item that was not acted upon.

## Findings:

*Equitable Evaluation of Faculty Teaching Performance.*

There are conflicts between various MSU faculty governance documents that concern the evaluation of faculty teaching performance and the influence that student surveys may have on these evaluations, particularly with respect to personnel decisions, including annual evaluations and salary decisions, as well as promotion and tenure decisions. These conflicts have created tensions between faculty, administrators and students, and this sets the stage for a variety of problems, including student frustration as well as faculty morale problems and even legal issues. Remediation will require better education of all parties, cooperation to harmonize operating policies and procedures, and vision to chart an aspirational path forward.

Some of the issue may be traced to IHL Policy 407.02 (Evaluating Teaching), which in turn governs MSU policy. IHL 407.02 states:

> *The method of annually evaluating the quality of teaching may employ multiple sources of data appropriate to the discipline. At a minimum, students enrolled in the course shall have the opportunity to provide written feedback about the faculty member's teaching effectiveness to the faculty member, department chair/head and academic dean.*

This policy recognizes that multiple sources of information may be used to evaluate the quality of teaching in a discipline-dependent manner, but then specifies as a minimum requirement that students must be given an opportunity to provide written feedback on a faculty member's teaching effectiveness. The policy leaves open what weight to ascribe student feedback in the evaluation process.

With IHL 407.02 as its basis, the current MSU AOP 13.15 (Evaluation of Teaching Performance) focuses on an instrument and the mechanism for collecting written feedback from students on the teaching effectiveness of faculty. The document notes that information collected from students should be combined with other measures of teaching performance where personnel decisions are being made or for assisting in faculty improvement. It goes on to specify that student evaluation shall not be the only criterion used in assessing teaching performance and lists several types of information that may be provided by faculty to support their teaching evaluation. AOP 13.15 does not specify or prioritize other measures of teaching performance that should be combined with information from student surveys and positioning this information at the end of the document may have the unfortunate effect of elevating student evaluations over these other measures.

The MSU Application for Promotion and/or Tenure as well as the Faculty Handbook also invoke student evaluations and teaching performance. The Application for Promotion and/or Tenure specifies (Sec. II.A.1) that applicants "must provide a summary statement of student evaluations," and then lists a variety of other materials that can be added to demonstrate teaching effectiveness. There is no specific guidance on what data the summary must contain. The Faculty Handbook notes (Sec. V.F. *Teaching*, pg. 30) that student evaluations are among the several types of documentary evidence that can be used to demonstrate teaching effectiveness.

Finally, information from student evaluations of teaching is not specifically required as a component in the annual activity reports that MSU faculty must file. However, "evaluations" are called out as a potential supporting documentation of teaching effectiveness, and many faculty members choose to incorporate all or part of the student evaluations they receive in their annual reports.

A large body of professional literature has accumulated to document the problems inherent in using student course evaluations in faculty personnel decisions (*as reviewed in* Kite et al. 2012, Linse 2017). Student responses reflecting clear racial and gender biases coupled with poor or misleading question design as well as over-interpretation of statistical analyses have led some institutions to go so far as to prohibit use of student course evaluations in any personnel decisions involving faculty. Others severely constrain how this information may be used in personnel decisions.

One hallmark of institutions actively seeking to elevate quality and innovation of instruction on their campuses is the adoption of an institutional standard for teaching excellence against which faculty and academic units may take the measure of their programs. These institutional standards are usually curated by the local equivalent of the MSU Center for Teaching and Learning and they are maintained in conjunction with a variety of tools and resources on which faculty can draw as they work to improve their teaching in light of these standards. At institutions that have made a deep cultural commitment to elevate teaching standards (e.g. Penn State University, University of Oregon, and University of Southern California) these institutional standards have been propagated to individual academic units where they have served to guide revisions to unit promotion and tenure guidelines. These revisions have frequently led academic units to specify as part of their promotion and tenure guidelines methods, tools and metrics beyond student course evaluations that will be used for assessment of faculty teaching effectiveness.

*Student Surveys.*

The current instrument used at MSU to collect student opinions about courses is called "Faculty Evaluation" and it comes in at least nine permutations distinguished by questions added to assess features that differentiate standard lecture courses from laboratory courses, studio courses and other specialty course formats encountered across campus. As specified in AOP 13.15, these "Faculty Evaluations" are managed by the Teaching Evaluation Committee. Despite faculty control of the instrument, there is widespread concern among many in the general faculty over use of the term *evaluation* with regard to these instruments and their results since students seldom have the necessary training or background to evaluate teaching quality. The fact that the current instrument assesses *class delivery* rather than *instructional quality* as perceived by students was even acknowledged in a November 2016 letter from the Executive Committee of the Holland Faculty Senate (copied to the Provost) responding to a request from the MSU Academic Deans for access to the written comments about courses and instructors collected via this instrument.

Additional specific concerns voiced to the task force about the current instrument for student evaluation of teaching included: 1) the frequency with which written comments include hostile ad hominem attacks on the instructor; 2) the inclusion of questions perceived to ask students to make judgements on pedagogical approaches for which they have no training; 3) a paucity of questions designed to probe classroom atmosphere and inclusivity issues; 4) perceived general biases and/or hostility toward instructors who are women, persons of color or representatives of other minority populations on campus; and 5) written comments describing faculty misbehavior that are only received long after the course is ended. Concerns were also voiced about the appropriateness of comparing averaged student evaluations scores for

individual faculty against composite averages reported for different departments and colleges, as well as for the university as a whole.

None of these problems is unique to MSU and many other institutions have already taken steps to address similar issues. Dr. Angela Linse, Executive Director and Associate Dean of the Scheyer Institute for Teaching Excellence at Penn State University has written a detailed and authoritative review of the subject (Linse 2017). In September 2019, the American Sociological Association, with endorsements from 17 additional academic associations, released a succinct and useful guiding statement on student assessments of teaching that drew recommendations from the work of numerous authorities, including Dr. Linse (ASA 2019). These two documents are useful guides to best practices for administrators, as well as promotion and tenure committees, seeking to incorporate student survey information into faculty teaching assessments.

Many other institutions have also moved in recent years to redesign the instruments they use to survey students about their classes in ways that focus more on student perceptions of features that characterize superior learning environments and excellence in teaching. When feasible, these surveys are administered at multiple points in the semester to provide faculty actionable feedback with which they can make course corrections during the term. When students see changes in response to this feedback, they are encouraged to take a more active approach in the entire process. Keeley et al. (2006, 2010, 2013) identified 28 instructor behaviors for which student survey responses in university psychology classes correlated well with teaching effectiveness. Questions focusing on student perception of behaviors associated with teaching excellence appear to provide reliable information about student learning and redesigned student surveys at other institutions have tended to feature questions that probe these behaviors with emphasis on how they impact the student's effort to learn.

There was considerable discussion in this task force regarding the precise makeup of a new student survey instrument. In the end, the group agreed upon the need for a new survey that focuses on "student reflection and experience" and has proposed a process for its creation in the recommendations (see Recommendation 4).

*Administration of Online Surveys.*

Student evaluations of courses have in the past been administered and collected in class where student participation could be monitored and encouraged by proctors. This is a costly exercise and peer institutions have increasingly moved to use of online surveys during the past two decades. The cost and logistics of administering hardcopy evaluations limited the frequency with which assessments could be made, typically to once per semester.

With the move to online surveys many institutions have reported an immediate drop in student response rates (Chapman and Joines 2017). This is a concern because low response rates are problematic for reliable statistical analyses of the collected data, although the impact may not be especially problematic against the backdrop of other issues already inherent to the data (Fike et al. 2010, Stanny and Arruda 2017). However, many institutions that that have shifted to using online student surveys provide recommendations to faculty for ways to promote increased participation of students in online surveys (see for example, Chapman and Joines 2017, University of Notre Dame 2019, University of Saskatchewan 2019).

*Related Discussion Item:*

At least one related item of committee discussion was not acted upon by the task force and is explained below. The group did not wish to include this as a recommendation; it is mentioned here as an item others may wish to pursue in the future. For this task force, the general feeling was that this was only tangentially related to our already weighty charge of evaluating instruction and that this group was not optimal for investigating or resolving this issue.

*Student Reports of Faculty Misbehavior*

Concern over comments written on student evaluations describing faculty misbehavior was a major driver for requests by Deans and Department Heads for access to these comments. Liability issues that could arise if the institution did not act on such reports likely drove the decision to revise IHL Policy 407.02 to require this access. As important as it is for the university to receive eye-witness reports of faculty misbehavior and/or malfeasance, no one would recommend end-of-the-semester student surveys as a vehicle for delivering such reports. Providing a single opportunity to report misbehavior only after the course has ended does not provide for timely intervention nor does it provide a means to report for students who drop a course, perhaps in response to faculty misbehavior.

One reason students use course evaluations to report faculty misbehavior is because this vehicle provides anonymity for the reporter. While anonymity is an important factor for encouraging students to report some of the most egregious types of faculty misbehavior, it also potentially provides cover for calumny by students. Students may also submit anonymous reports of faculty misbehavior at any time through the EthicsPoint portal, but the routing of reports from this portal may need to be reserved for the most egregious cases. While this task force has decided not to address this issue with a recommendation, it may merit further exploration by others to determine if a reporting system that is available on a continuous basis is advisable.

## Recommendations:

Charge 1: Develop proposed language for revising AOP 13.15, as necessary.

Charge 2: Determine whether adjustments are necessary to questions on the student evaluation of teaching effectiveness.

Charge 3: Strengthen language that encourages faculty teaching assessment beyond just using student surveys.

Charge 4: Develop plans to improve student surveys, increase participation and reduce faculty concerns about fairness and bias in student responses.

1. Commence a process to adopt an institutional "*Statement of Teaching Excellence at MSU*" that identifies shared values and expectations.

2. Require departments to use a variety of assessment measures to evaluate teaching and foster continual improvement. Assessment measures may include different types of peer evaluation and self-evaluation as well as student learning outcomes.

A. Provide educational programming and training (and/or other forms of information) regarding appropriate measures of teaching effectiveness to administrators, instructors and students.

B. Adopt optional standard documents and forms for a variety of peer and self-assessments to facilitate increased use of these assessments.

C. Consider requiring a basic, standard instructor self-assessment for each course taught.

D. End the practice of comparing an individual instructor's student survey results with departments, colleges or the university based on the research demonstrating that student surveys are not accurate measures of teaching effectiveness or of student learning, and that survey results are biased against faculty who are women and people of color.

3. Rename "Student Evaluations" to "Student Course Surveys" in order to de-emphasize the evaluative and comparative aspects of the results given the research that demonstrates that student surveys are not accurate measures of teaching effectiveness or of student learning. Change the way these data are used in order to minimize the impact of bias against faculty who are women and people of color in personnel decisions.

A. Provide a clear and succinct summary of the current research and data on Student Evaluations of Teaching (SETs) to all those responsible for personnel decisions regarding assessment of teaching. This summary to include:

1. Brief review of the current literature on bias in SETs

2. MSU Data on bias in SETs

3. Guidance on the use of SETs

B. Provide training for faculty and administrators in assessment of teaching and the appropriate use of information gathered from student surveys. Faculty engagement with and reflection upon student survey data will be considered a measure of teaching effectiveness.

C. Add explicit, targeted statement to the comments field in the student course survey to discourage inappropriate personal comments (see example below).

> Example from Iowa State University*: "Student evaluations of teaching play an important role in the review of faculty. Your opinions influence the review of instructors that takes place every year. Iowa State University recognizes that student evaluations of teaching are often influenced by students' unconscious and unintentional biases about the race and gender of the instructor. Women and instructors of color are systematically rated lower in their teaching evaluations than white men, even when there are no actual differences in the instruction or in what students have learned.*
>
> *As you fill out the course evaluation please keep this in mind and make an effort to resist stereotypes about professors. Focus on your opinions about the content of the course (the assignments, the textbook, the in-class material) and not unrelated matters (the instructor's appearance)."*

D. Develop an educational website and/or training program for faculty and students in regard to critical aspects of student course surveys including relevant research, constructive feedback and related diversity issues.

E. Distribute the proposed "Best Practices to Increase Student Response Rates to Online Course Surveys."

F. Consider adopting a standard, mid-semester student course survey that is for faculty who want to use it for formative feedback. In order to encourage use, the resulting data should only be provided directly to the faculty member for use in improving the course and student experience.

G. Change the student survey instrument to emphasize student reflection and experience and de-emphasize measurement of teaching effectiveness (see process outlined below).

4. Commence a process to create a new student survey instrument to focus upon student reflection and experience and in accordance with the values expressed in the proposed "*Statement of Teaching Excellence at MSU*."

<u>Suggested Process for Adoption of New Survey Instrument for Piloting in Fall 2021:</u>

Step 1: Presentations of relevant research and information in regard to student evaluations of teaching to be coordinated by the Center for Teaching and Learning in the spring of 2021.

Step 2: Creation of Student Survey Working Group (5-7 representatives) to assist in Development of New Survey instrument that focuses upon student experience and reflection.

Step 3: Feedback sessions with A. Teaching Evaluation Committee (current committee), B. Administrators, C. Teachers for survey input and D. Students. Sessions to be coordinated by the Center for Teaching and Learning in the spring of 2021.

Step 4: Presentation of New Survey to the existing Teaching Evaluation Committee for Review and Approval.

5. Institutionalize these measures into Promotion and Tenure guidelines and application forms housed in the Provost's Office, in each of the colleges, and in each department.

6. Institutionalize these measures in AOP 13.15 by refocusing the policy to create a more holistic, robust and equitable approach to evaluating teaching effectiveness (see proposed draft in Appendix).

## References:


American Sociological Association. “Statement on Student Evaluations of Teaching.” https://www.asanet.org/sites/default/files/asa_statement_on_student_evaluations_of_teaching_feb132020.pdf

Chapman, D. D., & Joines, J. A. (2017). Strategies for Increasing Response Rates for Online End-of-Course Evaluations. International Journal of Teaching and Learning in Higher Education, 29(1), 47-60.

Fike, David S., Denise J. Doyle, and Robert J. Connelly. 2010 "Online vs. Paper Evaluations of Faculty: When Less Is Just as Good." *Journal of Effective Teaching* 10.2: 42-54.

Holman, Mirya, Ellen Key and Rebecca Kreitzer. 2019. "Evidence of Bias in Standard Evaluations of Teaching." https://docs.google.com/document/d/14JiF-fT--F3Qaefjv2jMRFRWUS8TaaT9JjbYke1fgxE/edit

Keeley, Jared W., Dale Smith, and William Buskist. 2006. “The Teacher Behaviors Checklist: Factor Analysis of Its Utility for Evaluating Teaching.” Teaching of Psychology 33: 84-91. doi:10.1207/s15328023top3302_1

Keeley, Jared W., R. Michael Furr, and William Buskist. 2010. “Differentiating Psychology Students’ Perceptions of Teachers Using the Teacher Behavior Checklist.” Teaching of Psychology 37: 16-20. doi:10.1080/00986280903426282

Keeley, Jared W., Taylor English, Jessica Irons, and Amber Henslee. 2013. “Investigating Halo and Ceiling Effects in Student Evaluations of Instruction.” Educational and Psychological Measurement 73: 440-57. doi:10.1177/0013164412475300

Kite, Mary E., ed. 2012. Effective Evaluation of Teaching: A Guide for Faculty and Administrators. Washington, DC: Society for the Teaching of Psychology. http://teachpsych.org/ebooks/evals2012/index.php

Linse, Angela R. "Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees." Studies in Educational Evaluation 54 (2017): 94-106. https://www.sciencedirect.com/science/article/pii/S0191491X16300232

Stanny, Claudia and Arruda, James E. (2017). A Comparison of Student Evaluations of Teaching With Online and Paper-Based Administration. Scholarship of Teaching and Learning in Psychology. 3. 10.1037/stl0000087.

Stark, Phillip B. and Freishtat, Richard. (2014). An Evaluation of Course Evaluations. ScienceOpen Research. https://www.stat.berkeley.edu/~stark/Preprints/eval14.pdf

University of Notre Dame 2017 https://sites.nd.edu/kaneb/2017/04/17/improving-cif-response-rates/

University of Saskatchewan 2019 https://teaching.usask.ca/documents/seeq/online_course_evals.pdf

## Appendix 1: Proposed Revised AOP 13.15 Evaluation of Teaching Performance

### AOP 13.15: EVALUATION OF TEACHING PERFORMANCE

**PURPOSE**

The following policy guidelines have been adopted by the University to provide the faculty with a greater certainty of the procedure that will be used in the evaluation of teaching performance at Mississippi State University.

**POLICY/PROCEDURE**

Numerous measures of teaching performance can be used to assist in the process of faculty improvement and for personnel decisions. Personnel decisions in this case will include annual raises, annual evaluations, and promotion and/or tenure decisions.

Faculty members are expected to provide the department head and dean with information to support the evaluation of their teaching performance. A faculty member can choose among the following criteria to provide information to support evaluation of his or her teaching performance:

a) Peer evaluations (internal or external)
b) Self-evaluation or report
c) Classroom observation report
d) Student learning outcomes
e) Student course surveys
f) Faculty response to student course surveys
g) Faculty response to mid-term student course surveys
h) Scholarly research/publications/presentations related to teaching
i) Examples and/or analysis of course materials including course syllabi, assignments and exams
j) Teaching grants and awards
k) Additional student input in the form of letters, emails, faculty nominations, etc.
l) Curriculum development and innovation
m) Evidence of significant professional development in teaching
n) Additional evaluation materials.

Student course surveys will be administered uniformly across all courses each semester, but they shall not be the only criterion used to review teaching performance. Used alone, evaluation results may or may not provide accurate and appropriate information upon which to base judgments about teaching effectiveness. By themselves, student evaluations of teaching may indicate trends and provide faculty members with useful information about methods of instruction and practices. Used in combination with other types of information

about teaching performance, student course surveys can yield useful information about teaching effectiveness. Students will be informed of how the student course survey results will be used.

a) Student course surveys may be conducted using any mode(s) (e.g., electronic, paper) provided by and supported by the university.

b) The survey will investigate aspects of each of the following categories: (i) the course (ii) the instructor, and (iii) the method of delivery. The Teaching Evaluation Committee generally will be responsible for updating and changing the student course survey.

c) All procedures and processes for statistical reporting shall be developed and reviewed by the Teaching Evaluation Committee. The Teaching Evaluation Committee will consult with the Student Association.

The faculty member, along with their department head and dean or director, shall receive a copy of the statistical report and all comments for every evaluated class and section the individual teaches.

## REVIEW

This AOP will be reviewed every four years (or whenever circumstances require an earlier review) by the Associate Provost for Academic Affairs (APAA) with recommendationsfor revision presented to the Provost and Vice Presiden

## Appendix 2: Analysis of MSU's Course Survey Scores

The Office of Institutional Research and Effectiveness (OIRE) analyzed the results of the 11 common course survey questions and the global scores from spring 2017, 2018, 2019, and 2020 to determine the extent of bias against women and racial minorities in the data. The data included 12,752 course surveys for 2,823 non-duplicated faculty members.

In summary, the results indicate very little difference in scores (often at the hundredths decimal place). The combination of variables and controlling factors explained between 1.2% and 5.6% of the survey scores. Furthermore, the results indicated potential bias against males and minorities (Note: although there are variations in scores for specific questions, there was no significant difference in the global index scores for males or Black/African-American faculty):

- Males = 5/11 questions were significantly lower than females
- Racial minorities = 11/11 questions were significantly lower than white faculty
- Black/African American = 7/11 were significantly lower than white faculty

[Note: Racial minorities included American Indian, Asian, Black/African-American, Hispanic, Multi-racial, Pacific Islander/Alaskan Native, and International faculty. "Thick accent" is the number one student-reported concern with minority faculty, and the category for international faculty does not control sufficiently for an accent.]

OIRE used a linear regression to determine whether sex or race influenced a faculty member's score on a particular question when controlling for the following variables:

- Course level (lower division, upper division, masters, specialist, professional, doctoral)
- Whether the course was part of general education
- Whether the evaluation was completed online
- Whether the faculty member was tenured
- Whether the faculty member was tenure-track
- Whether the faculty member was a graduate student
- The faculty member's age

Table 1 (on the following page) provides the means for all faculty, males, minorities, and Black/African-American faculty. The highlighted cells indicate whether the difference in means is significant at the various probability levels. The adjusted $R^2$ value indicates what estimated percentage of the faculty members' scores could be determined by the combination of independent and control variables.

*Table 1. Means from student evaluations of teaching*

| | Mean | Std. Dev. | Male | Minority | Black/ African American | $R^2$ |
|---|---|---|---|---|---|---|
| The instructor created high expectations for the class. | 4.44 | 0.432 | 4.43 | 4.38 | 4.42 | 5.6 |
| The instructor conveyed the course content in an effective manner. | 4.25 | 0.610 | 4.23 | 4.14 | 4.19 | 4.3 |
| The instructor made the class interesting. | 4.18 | 0.639 | 4.18 | 4.07 | 4.17 | 4.7 |
| The instructor was enthusiastic about the subject matter. | 4.46 | 0.510 | 4.47 | 4.38 | 4.40 | 5.6 |
| The instructor was accessible outside of class time to respond to my questions or concerns. | 4.38 | 0.522 | 4.38 | 4.31 | 4.29 | 4.9 |
| I learned a great deal in this class. | 4.25 | 0.573 | 4.25 | 4.16 | 4.19 | 5.5 |
| The presentation of course content (lectures web materials and/or discussions etc.) helped me learn in this class. | 4.17 | 0.642 | 4.16 | 4.09 | 4.13 | 3.8 |
| The tests were fair. | 4.21 | 0.760 | 4.20 | 4.18 | 4.20 | 1.3 |
| The tests reflected material presented in lecture and/or assigned reading. | 4.32 | 0.714 | 4.31 | 4.29 | 4.29 | 1.2 |
| Tests and/or assignments were graded within a reasonable period of time. | 4.32 | 0.652 | 4.31 | 4.28 | 4.20 | 1.7 |
| I would recommend this instructor to other students if they wanted to learn this subject. | 4.28 | 0.648 | 4.29 | 4.18 | 4.22 | 4.4 |
| Global | 4.25 | 0.694 | 4.25 | 4.21 | 4.23 | 2.2 |
| Number of surveys analyzed | 12,752 | | 6,571 | 2,993 | 723 | |

significantly lower at the less than .001 level

significantly lower at the less than .01 level

significantly lower at the less than .05 level

## Appendix 3: Example of Recommended Definition or Statement of Teaching Excellence

From the *Definition of Teaching Excellence, University of Southern California*
http://cet.usc.edu/about/usc-definition-of-excellence-in-teaching/

### USC Definition of Excellence in Teaching

The University of Southern California is committed to excellence in teaching through the use of evidence-based, inclusive pedagogies that foster the knowledge, skills, relationships, and values necessary for students to succeed in a rapidly changing world. USC embraces an inclusive spirit that values the enrichment diversity brings to students' understanding, leading to greater opportunities to improve the lives of all people. It fosters a convergent spirit, teaching students to see problems and solutions from multiple viewpoints, to move fluidly across disciplines, and to work comfortably in disparate teams. And it cultivates an entrepreneurial spirit, empowering students to innovate and find creative approaches to solving complex problems. USC prepares students to navigate ambiguity, to utilize their intellectual curiosity to identify and realize opportunities, and to evolve into visionary leaders who seek impactful and ethical solutions for the local, national, and global challenges of our time.

1. **Respectful and Professional**
    a. Conveys commitment to learning through demonstrated effort in, and enthusiasm for, the teaching process
    b. Models and expects respectful and appropriate behavior in all professional interactions
    c. Develops professionalism in students through high expectations for mindful, ethical, responsible behavior
    d. Recognizes the power differential between professor and student, and acts with integrity toward students
    e. Fosters professional identity development through student use of discipline-specific customs and language
2. **Challenging and Supportive**
    a. Creates learning objectives and experiences that are challenging but attainable
    b. Models and fosters critical, analytical, and creative thinking
    c. Encourages student curiosity, exploration, and self-directed learning

d. Cultivates a belief that mistakes and failed experiments further knowledge and understanding
e. Fosters a mindset where growth is always possible, and ability is not fixed
f. Provides encouragement, positive reinforcement, and support
g. Guides students to university support services according to university policy

3. **Inclusive and Diverse**
    a. Creates an open environment conducive to intellectual risk-taking
    b. Includes students' strengths, experiences, and identities in the learning process
    c. Provides materials, cases, or applications that examine diverse experiences, perspectives, or populations
    d. Applies multiple techniques and strategies to reach all students in a culturally-responsive way
    e. Follows guidelines of Universal Design for Learning and accessibility best practices
4. **Relevant and Engaging**
    a. Uses content that is current, rigorous, and informed by theory, research, evidence, and context
    b. Uses active learning strategies to promote development of mastery
    c. Fosters transfer of learning and problem-solving skills to address real-world challenges
    d. Models and requires use of multiple media and technologies aligned with learning objectives and experiences
    e. Fosters student participation in academic discussions and fosters peer-to-peer collaboration, knowledge-sharing, and feedback
    f. Facilitates student engagement in inquiry and research
5. **Prepared and Purposeful**
    a. Uses instructional plan aligned with learning objectives that includes assessment of student prior knowledge, instruction followed by application, and shared reflection of what was learned
    b. Fosters self-regulation to help students to assess their own learning and adjust their strategies
    c. Manages learning effectively: plans activities, uses routines, and manages time, behavior, and participation
    d. Utilizes educational technologies (e. g., LMS) to provide students access to course materials, grades, and other feedback
6. **Fair and Equitable**
    a. Establishes clear expectations and learning objectives

b. Uses formative assessments to evaluate student progress, and summative assessments to evaluate mastery
c. Uses transparent assessment processes with clear criteria tied to learning objectives
d. Provides specific, regular, and timely feedback tied to performance criteria
e. Maintains reasonable course policies that are applied uniformly and fairly

**7. Evidence-Based**

a. Uses results from formative and summative peer and student teaching evaluations to inform teaching practice
b. Demonstrates effectiveness of instruction through measures of student mastery of learning objectives
c. Pursues continuous improvement of teaching and course design by applying research-based best practices

## Appendix 4: Best Practices to Increase Student Response Rates to Online Course Surveys

Online course surveys save money, lower staff workload, decrease the margin for error, preserve class time that would otherwise be spent on in-class surveys, and allow quick data turnaround. Still, going online is a big change from distributing paper surveys. As a faculty member, you naturally care about the feedback your students have to offer, and want to keep the accuracy, volume and quality of that feedback as high as possible. At institutions where student surveys are taken seriously by both the administration and the faculty, students feel that their feedback matters and respond accordingly. The relationship between student and faculty member is highly personal and often plays the biggest role in determining whether a student decides to submit a course survey.

**What can I do to maximize student response rates?**

1. Emphasize the significance of course surveys and let students know that their responses matter.
2. Make note of the survey period (typically 2-3 weeks before finals) in the course syllabus.
3. Encourage students to access the Class Climate survey tool through Canvas. (https://canvas.msstate.edu/)
4. Ask students to be on the lookout for email reminders to participate in Class Climate surveys and remind them to check their Junk mail folders if they don't receive reminders.
5. Monitor your Class Climate dashboard and communicate survey response rates to students during class.
6. Make the survey an assignment or exam that may be completed in Canvas during the survey period so that students are reminded to complete evaluations through pop-ups.
7. Set aside some time in a computer lab or during class, preferably at the beginning of the class period, when students can use laptops, tablets or cell phones to complete online surveys. (*Just as with paper evaluations, instructors must leave the classroom while students complete surveys.*)
8. Faculty may choose to set aside class time for surveys and tell students they will move forward with class once the overall response rate reaches 90-100% (depending on absences). The overall response rate can be projected from the Class Climate dashboard so students can follow the progress live. (*Just as with paper evaluations, instructors must leave the classroom while students complete evaluations.*)
9. Faculty may wish to consider incentives (e.g. points toward a participation grade) for students who submit surveys or, if not for individual students, provide incentives for entire classes that achieve a 100% response rate. The confirmation email each student receives after submitting their survey of a course may serve as proof that the student has participated. (*Research has failed to demonstrate introduction of additional bias when such micro-incentivization is used.*)

*Modified from recommendations made by Baylor University, Penn State University and University of Oregon.*

https://www.baylor.edu/irt/index.php?id=89690 1/1
https://www.schreyerinstitute.psu.edu/IncreaseSRTERespRate/
https://registrar.uoregon.edu/course-evaluations/response-rates-and-accuracy

**Appendix 5: Statement on Student Evaluations of Teaching by the American Sociological Association, September 2019.**

# Statement on Student Evaluations of Teaching

American Sociological Association
September 2019

Most faculty in North America are evaluated, in part, on their teaching effectiveness. This is typically measured with student evaluations of teaching (SETs), instruments that ask students to rate instructors on a series of mostly closed-ended items. Because these instruments are cheap, easy to implement, and provide a simple way to gather information, they are the most common method used to evaluate faculty teaching for hiring, tenure, promotion, contract renewal, and merit raises.

Despite the ubiquity of SETs, a growing body of evidence suggests that their use in personnel decisions is problematic. SETs are weakly related to other measures of teaching effectiveness and student learning (Boring, Ottoboni, and Stark 2016; Uttl, White, and Gonzalez 2017); they are used in statistically problematic ways (e.g., categorical measures are treated as interval, response rates are ignored, small differences are given undue weight, and distributions are not reported) (Boysen 2015; Stark and Freishtat 2014); and they can be influenced by course characteristics like time of day, subject, class size, and whether the course is required, all of which are unrelated to teaching effectiveness.

In addition, in both observational studies and experiments, SETs have been found to be biased against women and people of color (for recent reviews of the literature, see Basow and Martin 2012 and Spooren, Brockx, and Mortelmans 2015). For example, students rate women instructors lower than they rate men, even when they exhibit the same teaching behaviors (Boring, Ottoboni, and Stark 2016; MacNell, Driscol, and Hunt 2015), and students use stereotypically gendered language in how they evaluate their instructors (Mitchell and Martin 2018). The instrument design can also affect gender bias in evaluations; in an article in *American Sociological Review*, Rivera and Tilcsik (2019) find that the range of the rating scale (e.g., a 6-point scale versus a 10-point scale) can affect how women are evaluated relative to men in male-dominated fields. Further, Black and Asian faculty members are evaluated less positively than White faculty (Bavishi, Madera, and Hebl 2010; Reid 2010; Smith and Hawkins 2011), especially by students who are White men. Faculty ethnicity and gender also mediate how students rate instructor characteristics like leniency and warmth (Anderson and Smith 2005).

A scholarly consensus has emerged that using SETs as the primary measure of teaching effectiveness in faculty review processes can systematically disadvantage faculty from marginalized groups. This can be especially consequential for contingent faculty for whom a small difference in average scores can mean the difference between contract renewal and dismissal.

Given these limitations, the American Sociological Association, in collaboration with the scholarly societies listed below, encourages institutions to use evidence-based best practices for collecting and using student feedback about teaching (Barre 2015; Dennin et al. 2017; Linse 2017; Stark and Freishtat 2014). These include:

1. Questions on SETs should focus on student experiences, and the instruments should be framed as an opportunity for student feedback, rather than an opportunity for formal ratings of teaching effectiveness. For example, two universities – Augsburg University and University of North Carolina Asheville – recently revised and renamed their instruments to the "University Course Survey" and the "Student Feedback on Instruction Form," respectively, to emphasize that student feedback, while important, is not an evaluation of teaching effectiveness.

2. SETs should not be used as the only evidence of teaching effectiveness. Rather, when they are used, they should be part of a holistic assessment that includes peer observations, reviews of teaching materials, and instructor self-reflections. This holistic approach has been in wide use at teaching-focused institutions for many years and is becoming more common at research institutions as well. For example:
    - University of Oregon has undertaken a multi-year process to develop a holistic framework for assessing teaching effectiveness, including peer review, self-reflection, and student feedback. Extensive research and resources are available on the Office of the Provost website, including guidance on how to interpret SETs
    - University of Southern California has instituted peer review of teaching for faculty evaluation. Their Center for Excellence in Teaching provides resources for how to use peer review effectively and addresses common concerns.
    - University of California Irvine requires faculty to submit two types of evidence to document teaching effectiveness. In addition to SETs, faculty can submit a reflective teaching statement, peer evaluations of teaching, and other evidence like a Teaching Practices Inventory, developed by physicist Carl Weiman.
    - University of Nebraska Lincoln has articulated best practices for faculty evaluation that state, in part, "it is recommended that student evaluation scores should not be given undue weight in faculty evaluations, since these scores are easily manipulated and reflect many attitudes that extend beyond the successful accomplishment of the faculty member's teaching duties."
    - The University of Michigan's Center for Research on Teaching and Learning recommends that student ratings should never be used in isolation and should be part of a broader assessment of teaching effectiveness. They have developed resources that include a summary of research findings on SETs, a handout for students on how to make their feedback most helpful to instructors, and best practices for using SETs in personnel decisions.
    - Ryerson University has gone even further and is no longer using SETs for tenure or promotion decisions (Farr 2018). Instead, Ryerson asks faculty to compile a teaching dossier that includes a statement of teaching philosophy, evidence of curricular engagement, and self-reflections.
3. SETs should not be used to compare individual faculty members to each other or to a department average. As part of a holistic assessment, they can appropriately be used to document patterns in an instructor's feedback over time.
4. If quantitative scores are reported, they should include distributions, sample sizes, and response rates for each question on the instrument (Stark and Freishtat 2014). This provides an interpretative context for the scores (e.g., items with low response rates should be given little weight).
5. Evaluators (e.g., chairs, deans, hiring committees, tenure and promotion committees) should be trained in how to interpret and use SETs as part of a holistic assessment of teaching effectiveness (see Linse 2017 for specific guidance).

Gathering student feedback on their experiences in the classroom is an important part of student-centered teaching. This feedback can help instructors to refine their pedagogies and improve student learning in their courses. However, student feedback should not be used alone as a measure of teaching quality. If it is used in faculty evaluation processes, it should be considered as part of a holistic assessment of teaching effectiveness.

**Endorsements**

American Anthropological Association
American Dialect Society
American Folklore Society
American Historical Association
American Political Science Association
Archeological Institute of America
Association for Slavic, East European, and Eurasian Studies
Association for Theatre in Higher Education
Canadian Sociological Association
Dance Studies Association
International Center of Medieval Art
Korean American Communication Association
Latin American Studies Association
Middle East Studies Association
National Communication Association
National Council on Family Relations
National Council on Public History
Rhetoric Society of America
Society for Cinema and Media Studies
Society for Classical Studies
Society for Personality and Social Psychology
Society of Architectural Historians
Sociologists for Women in Society

**References**


Anderson, Kristin J., and Gabriel Smith. 2005. "Students' Preconceptions of Professors: Benefits and Barriers According to Ethnicity and Gender." *Hispanic Journal of Behavioral Sciences* 27(2):184-20.

Barre, Elizabeth. 2015. "Student Ratings of Instruction: A Literature Review." Reflections on Teaching and Learning: The CTE Blog. Rice University Center for Teaching Excellence. Retrieved from https://cte.rice.edu/blogarchive/2015/02/01/studentratings.

Basow, Susan A., and Julie L. Martin. 2012. "Bias in Student Evaluations." Pp. 40-49 in *Effective Evaluation of Teaching: A Guide for Faculty and Administrators*, edited by Mary E. Kite. Washington, DC: Society for the Teaching of Psychology. Retrieved from http://teachpsych.org/ebooks/evals2012/index.php.

Bavishi, Anish, Juan M. Madera, and Michelle R. Hebl. 2010. "The Effect of Professor Ethnicity and Gender on Student Evaluations: Judged Before Met." *Journal of Diversity in Higher Education* 3(4):245-256.

Boring, Anne, Kellie Ottoboni, and Philip B. Stark. 2016. "Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness." *ScienceOpen Research* (DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1).

Boysen, Guy A. 2015. "Significant Interpretation of Small Mean Differences in Student Evaluations of Teaching Despite Explicit Warning to Avoid Overinterpretation." *Scholarship of Teaching and Learning in Psychology* 1(2):150-162.

Dennin, Michael, Zachary D. Schultz, Andrew Feig, Noah Finkelstein, Andrea Follmer Greenhoot, Michel Hildreth, Adama K. Leibovich, James D. Martin, Mark B. Moldwin, Diane K. O'Dowd, Lynmarie A. Posey, Tobin L. Smith, and Emily R. Miller. 2017. "Aligning Practice to Policies: Changing the Culture to Recognize and Reward Teaching at Research Institutions." *CBE—Life Sciences Education* 16(5):1-8.

Farr, Moira. 2018. "Arbitration Decision on Student Evaluations of Teaching Applauded by Faculty." *University Affairs* August 28. Retrieved from https://www.universityaffairs.ca/news/news-article/arbitration-decision-on-student-evaluations-of-teaching-applauded-by-faculty/.

Linse, Angela R. 2017. "Interpreting and Using Student Ratings Data: Guidance for Faculty Serving as Administrators and on Evaluation Committees." *Students in Educational Evaluation* 54:94-106.

MacNell, Lillian, Adam Driscoll, and Andrea N. Hunt. 2015. "What's in a Name: Exposing Gender Bias in Student Ratings of Teaching." *Innovative Higher Education* 40(4):291-303.

Mitchell, Kristina M. W., and Jonathan Martin. 2018. "Gender Bias in Student Evaluations."

*PS: Political Science and Politics* 51(3):648-652.

Reid, Landon 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.

Rivera, Lauren A., and András Tilcsik. 2019. "Scaling Down Inequality: Rating Scales, Gender Bias, and the Architecture of Evaluation." *American Sociological Review* 84(2):248–274.

Smith, Bettye P., and Billy Hawkins. 2011. "Examining Student Evaluations of Black College Faculty: Does Race Matter?" *The Journal of Negro Education* 80(2):149-162.

Spooren, Pieter, Bert Brockx, and Dimitri Mortelmans. 2013. "On the Validity of Student Evaluation of Teaching: The State of the Art." *Review of Educational Research* 83(4):598–642.

Stark, Philip B., and Richard Freishtat. 2014. "An Evaluation of Course Evaluations." *ScienceOpen Research* (DOI: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1).

Uttl, Bob, Carmela A. White, and Daniela Wong Gonzalez. 2017. "Meta-Analysis of Faculty's Teaching Effectiveness: Student Evaluation of Teaching Ratings and Student Learning Are Not Related." *Studies in Educational Evaluation* 54(1):22-42.

**Additional Resources**

Association of American Universities. n.d. *Aligning Practice to Policies: Changing the Culture to Recognize and Reward Teaching at Research Universities.* Retrieved from https://www.aau.edu/sites/default/files/AAU-Files/STEM-Education-Initiative/Aligning-Practice-To-Policies-Digital.pdf

American Educational Research Association. 2013. *Rethinking Faculty Evaluation: AERA Report and Recommendations on Evaluating Education Research, Scholarship, and Teaching in Postsecondary Education.* Retrieved from http://www.aera.net/Portals/38/docs/Education_Research_and_Research_Policy/RethinkingFacultyEval_R4.pdf.

Center for the Integration of Research, Teaching, and Learning (CIRTL). 2018. Local CIRTL Program Evaluation Resource Guide. Michigan State University. Retrieved from https://www.cirtl.net/resources/708.

Kite, Mary E., ed. 2012. *Effective Evaluation of Teaching: A Guide for Faculty and Administrators*. Washington, DC: Society for the Teaching of Psychology. Retrieved from http://teachpsych.org/ebooks/evals2012/index.php.

Stark, Philip B. 2018. "Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness." Presentation given as part of the Teaching Assessment Working Group Speaker Series, Valuing Teaching: Challenges and Strategies to Value Effective Teaching. Simon Fraser University. Retrieved from https://www.youtube.com/watch?v=5haOjlfJDb8&feature=youtu.be.

Van Valey, Thomas L., ed. 2011. *Peer Review of Teaching: Lessons from and for Departments of Sociology*. Washington, DC: American Sociological Association.

Vasey, Craig, and Linda Carroll. 2016. "How Do We Evaluate Teaching? Findings from a Survey of Faculty Members." *Academe*, May-June. Retrieved from https://www.aaup.org/article/how-do-we-evaluate-teaching.

**Appendix 6: Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees by Angela R. Linse**

Studies in Educational Evaluation 54 (2017) 94–106

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

# Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees

Angela R. Linse

*The Pennsylvania State University, United States*

ARTICLEINFO

*Article history:*
Received 4 February 2016
Received in revised form 3 December 2016
Accepted 6 December 2016
Available online 20 February 2017

*Keywords:*
Faculty evaluation
Student ratings
Personnel evaluation
Evaluation usage
Evaluators

ABSTRACT

This article is about the accurate interpretation of student ratings data and the appropriate use of that data to evaluate faculty. Its aim is to make recommendations for use and interpretation based on more than 80 years of student ratings research. As more colleges and universities use student ratings data to guide personnel decisions, it is critical that administrators and faculty evaluators have access to research-based information about their use and interpretation.

The article begins with an overview of common views and misconceptions about student ratings, followed by clarification of what student ratings are and are not. Next are two sections that provide advice for two audiences—administrators and faculty evaluators—to help them accurately, responsibly, and appropriately use and interpret student ratings data. A list of administrator questions is followed by a list of advice for faculty responsible for evaluating other faculty members' records.



## 1. The problem: misinterpretation and misuse of student ratings data

> *Steadily accumulating evidence of the misuse or overuse of ratings data . . . and the perennial debate in the press concerning the validity of student ratings . . . do not invalidate the potential of ratings data as useful information about teaching performance.* (Theall & Franklin, 2000, p. 95)

Student ratings instruments have been around since the 1920s (Marsh, 1987; Remmers, 1933; Remmers & Brandenburg, 1927). I use the term student ratings to refer to surveys administered by colleges and universities directly to enrolled students under controlled circumstances, typically near the end of an academic term. These surveys are also referred to as student evaluations of teaching (SETs), student ratings of instruction (SRIs), teaching evaluations, and course evaluations.

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). 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; Miller & Seldin, 2014; Seldin, 1999). In addition to serving as a source of feedback for instructional improvement, at most institutions student ratings are also used in personnel decisions such as annual reviews, merit raises, tenure and promotion, post-tenure review, and for hiring and re-appointment of "tenure exempt" faculty.[1] The challenge of appropriate use of student ratings data will be with us as long as we continue to use them.

The purpose of this article is to make recommendations about some of the most common misuses of student ratings data in the faculty evaluation process, in a format that can be easily shared. But first, I briefly justify the need for this article by reviewing the common misconceptions of student ratings and faculty concerns about student ratings as represented in the academic press. Next, I suggest that the vast body of research literature on student ratings generally refutes the misconceptions, but that this literature is not widely known or accessed by faculty and administrators. The paper ends with two sections of concise and candid guidance for two groups based on the challenges they face in using student ratings for evaluation: 1) administrators who must be able to accurately answer faculty questions about how their student ratings will be used and interpreted; and 2) faculty responsible for evaluating other faculty members' dossiers. These guides fill an important gap

[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.

*E-mail address:* arlinse@gmail.com (A.R. Linse).

http://dx.doi.org/10.1016/j.stueduc.2016.12.004

in the faculty evaluation literature created by a lack of formal training in use and interpretation of student ratings data, which leaves faculty and administrators to gather information based on their own experiences and the easily accessible academic press.

This article does not provide yet another research study or more empirical evidence that student ratings instruments are effective for gathering student feedback. Neither is this article intended to dispel myths about student ratings, nor provide a comprehensive overview of the vast student ratings research literature. Numerous other authors provide reviews and summaries of the research literature (Benton & Cashin, 2011; Benton & Li, 2015; Berk, 2005, 2013; Cashin, 1999, 2003). Readers interested in how to create a valid and reliable faculty evaluation system should consult Arreola (2007), Berk (2006), Braskamp, Brandenburg, & Ory (1984), Cashin (1996) and Hativa (2013a). To develop an in-depth understanding of the history and leaders of student ratings research, readers are directed to the works of 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, Braskamp, & Pieper, 1980).

## 2. Common views about student ratings

This article was, in part, prompted by the misinformation about student ratings that is easily accessible on the web and which is widely shared among faculty (Barre, 2015). Every few years, clusters of stories appear in the academic press that claim to have found fatal flaws in student ratings of teaching (e.g., Berrett, 2015a; Burt, 2015; Flaherty, 2016a). These stories are occasionally picked up by other news organizations (e.g., Barlow, 2015; Harvard Business Review, 2014; National Public Radio, 2015; Schuman, 2014). These stories raise fear among faculty members that they are, or will be, subject to unfair use of student ratings. Sensational headlines merge with a steady stream of stories that ensure anxieties about student ratings persist among the faculty.

Since 2007, the two academic news organizations most widely read by faculty in the U.S., *The Chronicle of Higher Education* and *Inside Higher Education*, have published more than 50 stories about (or implicating) student ratings. Of these, almost 65% percent are negative, while only about 10% include both positive and negative comments about student ratings. Many of these stories are opinion pieces or essays that do not cite research to support their claims (e.g., Basu, 2011; Edwards, 2012; Epstein, 2010; Eubanks, 2011; Fant, 2010; Haynie, 2010; Inchausti, 2014; Jafar, 2012; Moriarty, 2009; Warner, 2012a, 2012b). Others report on studies that have not been peer reviewed or published (e.g., Berrett, 2015b; Fischman, 2010; Glenn, 2007, 2010; Pettit, 2016; Zaino, 2015) or that are of limited applicability because they examine student ratings in a single discipline or from a narrow (and not necessarily representative) segment of the student population (Breslow, 2007; Glenn, 2011; Hamermesh, 2011; Heggen, 2008; Powers, 2007). Less than 25% of the studies are positive or include useful advice (e.g., Aragon, 2013; Dean Dad, 2007, 2010; Miller, 2010; Perlmutter, 2011; Sprague, 2016; Warner, 2012a, 2012b; Weir, 2010). Almost none of the 50 stories note that the issues raised were identified and examined long ago by student ratings researchers.

The most sensational headlines suggest that student ratings have finally been recognized as hopelessly flawed and/or predict their imminent demise (see above citations), but they do reflect the concerns of faculty, including that:

- Student ratings are the sole measure of teaching
- Other faculty manipulate students to achieve higher ratings
- Students are biased against certain faculty members (and no one will notice)
- Ratings do not reflect use of effective teaching methods
- Correlations with other variables make the ratings invalid or unreliable
- Online response rates are too low to be representative
- Students do not take the ratings seriously, lie, or are overly critical
- Evaluators focus on rare or negative ratings and do not know what normal variation is acceptable

Based on the regular appearance of articles questioning the value and use of student ratings and suggesting that they are universally reviled by faculty (e.g., Bernhard, 2015; Patton, 2015), two conclusions can be drawn. First, concerns important to the faculty about the use of student ratings have not been sufficiently addressed. Second, what we know about student ratings from the research literature is not reaching faculty or administrators. Faculty and administrators are largely unaware of the vast research literature, even though it is the most researched topic in higher education (Berk, 2013; Seldin, 1999) and the research literature has accumulated for more than 80 years (Cashin, 1999; Ory, 2001; Theall & Franklin, 1990, 2001).

## 3. What student ratings are and are not

> *The students' satisfaction with, or perception of, learning is related to the evaluations they give.* (Clayson, 2009, p. 26)

Before advancing to the primary sections of this article, Questions Asked by Administrators and Guidelines for Faculty, it is important to clarify what student ratings are and are not.

*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[2] (Abrami, 2001; Arreola, 2007; Hativa, 2013a).Data are typically collected systematically from enrolled students who have experienced the learning environment created by the faculty member. Most student ratings instruments include a series of items with rating scales that ask about students' perceptions in terms of quality, agreement, importance, frequency, or likelihood. The scales are typically linear, ordinal, and divided into five to seven categories. Some instruments use numerical rating scales anchored at each end with "highest rating" and "lowest rating."

*Student ratings are not faculty evaluations.*

Student ratings researchers are clear to differentiate between the producers of the data (students) and the users of the data (faculty and administrators) for both improvement and evaluative purposes. That many faculty view student ratings as evaluations likely stems from the names colleges and universities assign to their ratings instruments, e.g., Student Evaluations of Teaching, Course Evaluations).

*Student Ratings Are Not Measures of Student Learning.*

Student ratings have never been intended to serve as a proxy for learning. Confusion over this may result from student ratings

[2] Student ratings administered by a college or university are not the same as publicly available ratings websites, such as ratemyprofessors.com. Such sites are open to anyone, not solely to enrolled students, and they rely entirely on students motivated to visit the site.

research that has demonstrated a low to moderate positive correlation between students' ratings and their grades or expected grades (Abrami, 2001; Abrami, Dickens, Perry, & Leventhal, 1980; Benton & Li, 2015; Eiszler, 2002; Feldman, 1976; Greenwald & Gillmore, 1997; Stumpf & Freedman, 1979). Even though grades are supposed to reflect student learning, a simple correlation between grades and student ratings does not demonstrate causality, i.e., that high grades result in high ratings. Faculty who teach well, have grading practices that are accurate reflections of students' learning, and have grade distributions with a peak near the high end of the grading scale, may receive higher ratings—and deservedly so.

*Student Ratings Are Here to Stay.*

Given the utility of student ratings in academic decision making, student ratings are unlikely to be eliminated any time soon (Benton & Cashin, 2011; Franklin, 2001; Kulik, 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).

## 4. Ensuring appropriate interpretation and use of student ratings data

> *Not only can students provide data about the effects that instruction has had on them, but they also have an excellent opportunity to observe what the teacher does and what the course requires. Thus student reports of instruction have commonly been used as a source of data, not only for research, but also to improve teaching and to evaluate teaching for personnel decisions* (McKeachie, 1990, p. 194)

Faculty rotate on to and off of review committees and faculty move into new administrative roles that require evaluation of other faculty. Yet, faculty in evaluative roles are rarely, if ever, provided guidelines for interpreting others' student ratings. Without research-based guidance these faculty and administrators are likely to view other faculty members' student ratings through the lens of their own experience. New administrators eventually may see a wide range of student ratings and develop an understanding of the variability across courses and individuals. However, faculty on review committees may only see the ratings for a handful or two of faculty per year.

In order for faculty administrators and members of faculty review committees to accurately, responsibly, and appropriately interpret data derived from student ratings of instruction, they need access to recommendations founded in the research literature.

Many of the unresolved faculty concerns listed above are addressed in the two sections below, Questions Asked by Administrators and Guidelines for Faculty. Only those concerns that are implicated in the use of student ratings for the evaluation of faculty are discussed.

## 5. 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, e.g., in pre-tenure reviews or reappointments. Below are some of the most common questions asked by administrators and faculty. This section reflects common faculty misconceptions of student ratings, not just those held by faculty who receive low ratings or who are unhappy with their results.

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.

### *5.1. 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, 2013a, 2013b).

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!

### *5.2. 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. This needs 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 do improve (Boice, 2001; Brinko, 1991; Geis, 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 self-selected to seek 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 improvements and address students' concerns. Many centers also provide consultation services to administrators who are seeking advice about how to mentor faculty within their units.

### 5.3. *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 (Gigliotti & Buchtel, 1990; Marsh & Roche, 1997), 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 & Gillmore, 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 & Li, 2015).

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.

### 5.4. *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, while student ratings are not necessarily a "popularity contest," 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, d'Apollonia, & Cohen, 1990; Aleamoni, 1999; d'Appolonia & Abrami, 1997; Feldman, 1989; Marsh, 1982b, 1984; 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 (Berk, 2005, 2013; McKeachie, 1997) and they are highly correlated with other measures of teaching effectiveness (Abrami et al., 1990; Berk, 2013). Unfortunately, this may not change minds because statistical validity is not really the concern.

When faculty question the validity of students ratings, they are typically not concerned about the statistical validity or reliability of the ratings instrument, but instead they are concerned whether their ratings will be used against them. This provides an opportunity to talk about many of the issues discussed in this article.

If neither of these strategies works, be honest that student ratings are unlikely to become obsolete any time soon, no matter what the latest headlines say. Student ratings have been around

since the 1920s and they 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 teaching and the learning environment for students. Specifically, instructors who want to increase their ratings should focus their efforts on improving the learning environment for students through "communication, motivational, and rapport-building skills" (IDEA Research Note 1, 2003). Campus teaching and learning centers have many resources and strategies to help faculty with these attributes of effective teaching.

### 5.5. 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 to solely reflect of that 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 identity, political, religious, etc.). While these biases definitely exist, the research indicates that the biases rarely, if ever, fully explain the student ratings results for a faculty member who consistently receives ratings clustered 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. These biases are more difficult to detect. Over time, a growing body of research has been able to document 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, McLoughlin, & Garcia, 1999; Basow, 1995; Centra & Gaubatz, 2000; Hancock, Shannon, & Trentham, 1993; Sinclair & Kunda, 2000). While recent stories in the academic press (e.g., Flaherty, 2016b) have generated a lot of attention, the articles cited (Braga, Paccagnella, & Pellizzari, 2014; MacNell, Driscoll, & Hunt, 2015) have methodological issues, and significantly overstate the case (Ryalls, Benton, Barr, & Li, 2016).

The research on gender bias has a longer history than does the research on bias due to race, ethnicity, or culture, in part because faculty with non-majority attributes are still a relatively small percentage of the faculty. However, the number of studies is increasing and evidence is mounting that such biases also 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, again, at this point the bias is not sufficiently strong or widespread to explain consistently low ratings across all courses for a faculty member.

### 5.6. 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 ThePennsylvania 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,[3] which processes student ratings from hundreds of institutions, reports no difference in online ratings (Webster, Benton, & Gross, 2010) as do numerous other studies (Dommeyer, Baum, Hanna, & Chapman, 2004; McGhee & Lowell, 2003; Stowell, Addison, & Smith, 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, a note should be included in the faculty member's dossier, preferably in the department chair's statement.

### 5.7. 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

[3] 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.

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, Austin, Eddy, & Beach, 2006).

### 5.8. *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 (Fairweather, 2002; Glassick, Huber, & Maeroff, 1997; 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."

### 5.9. *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. Response rates for online administration tend to fall by 25–30% (Benton, Webster, Gross, & Pallett, 2010; Hativa, 2013a; Johnson, 2003; Nulty, 2008; Sorenson & Reiner, 2003), but may again increase as students no longer expect paper student ratings and mobile versions again 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 his/her efforts to improve the course, noting that their feedback will likely benefit future students, and multiple reminders from the faculty. Many online systems are programed to provide automatic reminders when a student has unrated courses. 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 grade rewards amount to bribery for higher ratings. Two practices that are extremely successful include granting students early access to grades or 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 (https://sites.psu.edu/academicaffairs/files/2016/09/srte_statement-248pj9j.pdf) "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." 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.

### 5.10. *How do I respond to faculty members 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, Ardalan, Coppage, & Crouch, 2007; Dommeyer et al., 2004; Hardy, 2003; Venette, Sellnow, & McIntyre, 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.

## 6. Guidelines for faculty who use student ratings data to evaluate other faculty

> *As the importance of teaching evaluation rises, we must examine means of evaluation to ensure that we are furthering—not hindering—teaching excellence.* (Miller & Seldin, 2014, p.1)

### 6.1. 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.

### 6.2. 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, schools, colleges) combine a single 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. 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 difficulty of teaching a particular type of course or context should be avoided (e.g., adding a 0.2 points for teaching a course larger than 50).

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 |

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).

### 6.3. 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 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

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 |

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 will differ 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 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 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 the low ratings directly to the faculty member's teaching.

The argument for not over-interpreting relatively small differences in average ratings is supported by the 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 (Arreola, 2007; Feldman, 2007; Hativa, 2013b). Bias due to gender, race, ethnicity, or culture is addressed in the previous section under the question about student bias.

### 6.4. *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 an effort to improve.

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.

### 6.5. *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 it also makes the mean (average) not the best measure of central tendency for the distribution. 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.

*6.6. 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, Kimmel, & Kromrey, 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).

*6.7. 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.

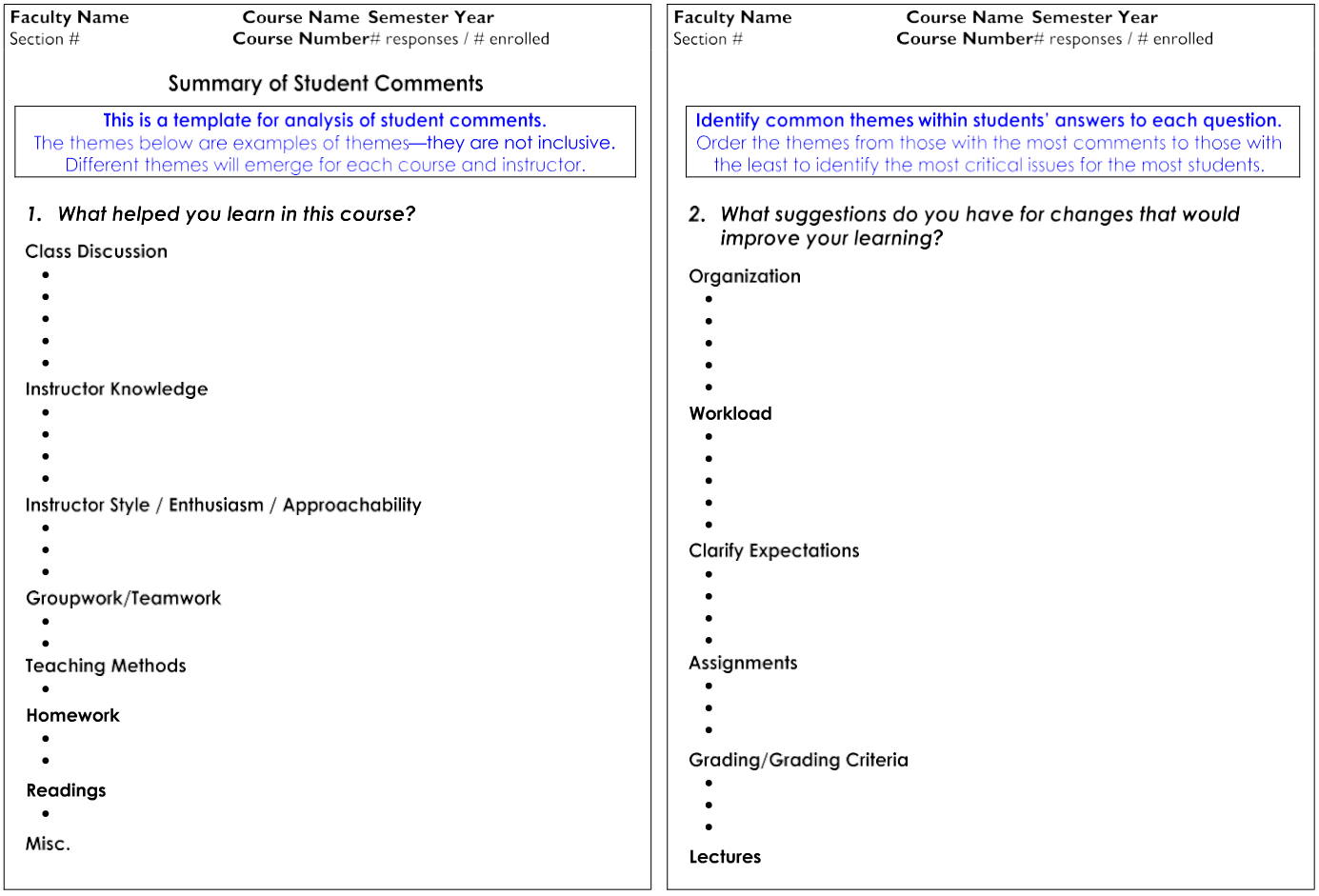


Figure 1. Sample format for a thematic analysis of students' written comments.

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, Ory, & Pieper, 1981; Marincovich, 1999; Ory et al., 1980). But too often, faculty and administrators seem to focus their attention on rare comments, possibly because they 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 awarded 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. 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 to not over-emphasize isolated comments, whether positive or negative.

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.

### 6.8. 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 to be 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.

## 7. Closing remarks

In sum, this article makes a number of points. 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 obtained instead 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 article 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? *Journal 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.

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.

Aragon, J. (2013). Re-evaluating My Relationship with Student Evaluations. *Inside Higher Education*, March 3, 2013. Retrieved from https://www.insidehighered.com/blogs/university-venus/re-evaluating-my-relationship-student-evaluations.

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.

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

Barlow, A. (2015). Culling the Iowa Faculty, Academe Blog, American Association of University Professors, April 21, 2015. Retrieved from https://academeblog.org/2015/04/21/culling-the-iowa-faculty/.

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. Retrieved from 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.

Basu, K. (2011). Socratic backfire? *Inside Higher Education*, October 31, 2011. Retrieved from https://www.insidehighered.com/news/2011/10/31/after-student-complaints-utah-professor-denied-job.

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. Retrieved from http://ideaedu.org/wp-content/uploads/2014/11/idea-paper_50. pdf.

Benton, S. L., & Li, D. (2015). *Response to A Better Way to Evaluate Undergraduate Teaching, IDEA Editorial Note #1, IDEA Center*. Retrieved from http://ideaedu.org/research-and-papers/editorial-notes/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. Retrieved from http://ideaedu.org/wp-content/uploads/2014/11/techreport-16.pdf.

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.

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. (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. (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.

Bernhard, M. (2015). Everyone complains about evaluations. *The Chronicle of Higher Education*, June 15, 2015. Retrieved from http://www.chronicle.com/article/Everyone-Complains-About/230885.

Berrett, D. (2015a). Can the student course evaluation be redeemed? *The Chronicle of Higher Education*, November 29, 2015. Retrieved from http://www.chronicle.com/article/Can-the-Student-Course/234369.

Berrett, D. (2015b). Scholars take aim at student evaluations' 'air of objectivity.' *The Chronicle of Higher Education*, September 18, 2015. Retrieved from http://www.chronicle.com/article/Scholars-Take-Aim-at-Student/148859.

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.

Braga, M., Paccagnella, M., & Pellizzari, 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.

Breslow, J. M. (2007). A glance at the September issue of social science quarterly. *The Chronicle of Higher Education*, October 02, 2007. Retrieved from http://www.chronicle.com/article/Adjusting-for-Bias-in-Student/1591.

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

Burt, S. (2015). Why not get rid of student evaluations? The answer requires us to think about power, *Slate*, May 15, 2015. Retrieved from http://www.slate.com/articles/life/education/2015/05/a_defense_of_student_evaluations_they_re_biased_misleading_and_extremely.html.

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

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

Cashin, W. E. (1999). Student ratings of teaching: uses and misuses. In P. Seldin (Ed.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 25–44). Bolton, Massachusetts: Anker.

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

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.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31*(1), 16–29.

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.

"Dean Dad" (pseudonym) (2007). Reading Evaluations. *Inside Higher Education*, December 5, 2007. Retrieved from https://www.insidehighered.com/blogs/confessions_of_a_community_college_dean/reading_evaluations.

"Dean Dad" (pseudonym) (2010). How to Read Student Evaluations. *Inside Higher Education*, December 16, 2010. Retrieved from https://www.insidehighered.com/blogs/confessions_of_a_community_college_dean/how_to_read_student_evaluations.

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

Edwards, T. (2012). The inherent unreliability of student evaluations. *The Chronicle of Higher Education*, March 21, 2012. Retrieved from http://www.chronicle.com/article/The-Inherent-Unreliability-of/131232.

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

Epstein, J. (2010). Grades on the rise. *Inside Higher Education*, March 5, 2010. Retrieved from https://www.insidehighered.com/news/2010/03/05/grades.

Eubanks, P. (2011). Why we inflate grades. *Inside Higher Education*, August 9, 2011. Retrieved from https://www.insidehighered.com/views/2011/08/09/essay_on_why_faculty_members_participate_in_grade_inflation.

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

Fant, G. C., Jr. (2010). Tricks for boosting student evaluations. *The Chronicle of Higher Education*, March 24, 2010. Retrieved from http://www.chronicle.com/blogs/onhiring/tricks-for-boosting-student-evaluations/22033.

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

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

Fischman, J. (2010). Students lie on course evaluations, study finds. *The Chronicle of Higher Education*, December 13, 2010. Retrieved from http://www.chronicle.com/blogs/ticker/students-lie-on-course-evaluations-study-shows/29079.

Flaherty, C. (2016a). Bias against female instructors. *Inside Higher Education*, January 11, 2016. Retrieved from https://www.insidehighered.com/news/2016/01/11/new-analysis-offers-more-evidence-against-student-evaluations-teaching.

Flaherty, C. (2016b). Zero correlation between evaluations and learning. *Inside Higher Education*, September 21, 2016. Retrieved from https://www.insidehighered.com/news/2016/09/21/new-study-could-be-another-nail-coffin-validity-student-evaluations-teaching.

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).

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, Illinois, April 7, 1991. ERIC # ED 349 318.

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, Louisiana, April 7, 1994.

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), 26–27 July 2, 2007.

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

Glenn, D. (2007). Method of using student evaluations to assess professors is flawed but fixable, 2 scholars say. *The Chronicle of Higher Education*, May 29, 2007. Retrieved from http://www.chronicle.com/article/Method-of-Using-Student/31696.

Glenn, D. (2010). 2 studies shed new light on the meaning of course evaluations. *The Chronicle of Higher Education*, December 19, 2010. Retrieved from http://www.chronicle.com/article/2-Studies-Shed-New-Light-on/125745/.

Glenn, D. (2011). One measure of a professor: students' grades in later courses. *The Chronicle of Higher Education*, January 9, 2011. Retrieved from http://www.chronicle.com/article/One-Measure-of-a-Professor-/125867.

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. (2016). *Women in the academy: Learning from our diverse career pathways*. Lanham, Maryland: Lexington.
Hamermesh, D. S. (2011). Beauty pays. *Inside Higher Education*, August 15, 2011. Retrieved from https://www.insidehighered.com/views/2011/08/15/beauty-pays.
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.
Harvard Business Review (2014). Better teachers receive worse student evaluations, September 2014. Retrieved from https://hbr.org/2014/09/better-teachers-receive-worse-student-evaluations/.
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.
Haynie, A. (2010). Motherhood after tenure: Evaluation time. *Inside Higher Education*, February 4, 2010. Retrieved from https://www.insidehighered.com/blogs/mama_phd/motherhood_after_tenure_evaluation_time.
Heggen, J. (2008). Evaluating faculty quality, randomly. *Inside Higher Education*, July 11, 2008. Retrieved from https://www.insidehighered.com/news/2008/07/11/evaluation.
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.
Inchausti, R. (2014). Scrap those old evaluation questions: Use these instead. *The Chronicle of Higher Education*, June 10, 2013. Retrieved from http://www.chronicle.com/blogs/letters/scrap-those-old-evaluation-questions-use-these-instead/.
Jafar, A. (2012). Warning: Reading student evaluations can make you crazy. *Inside Higher Education*, June 24, 2012. Retrieved from https://www.insidehighered.com/blogs/university-venus/warning-reading-student-evaluations-can-make-you-crazy.
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 G. Gutiérrez y Muhs, Y. F. Niemann, C. G. González, A. P. Harris (Eds.), Presumed Incompetent: The Intersections of Race and Class for Women in Academia, (pp. 164–185). Boulder, Colorado: Utah State University Press, an imprint of Colorado University Press.
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.
Lewis, K. G. (2001). Making sense of student written comments. *New Directions for Teaching and Learning, 87*, 25–32.
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 P. Seldin (Ed.), *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions* (pp. 45–69). Bolton Massachusetts: Anker.
Marsh, H. W. (1980). Research on students' evaluations of teaching effectiveness. *Instructional Evaluation, 4*(5), 5–13.
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, 74*, 264–279.
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. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253–369.
Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–384).New York New York: Springer.
Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In C. J. Smart (Ed.), *Higher education: handbook of theory and research*, Volume 8 (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. (1979). Student ratings of faculty: A reprise. *Academe, 65*(6), 384–397.
McKeachie, W. J. (1990). Research on college teaching: the historical background. *Journal of Educational Psychology, 82*(2), 189–200.
McKeachie, W. J. (1997). Student ratings: the validity of use. *American Psychologist, 52*(11), 1218–1225.
Miller, J. E., & Seldin, P. (2014). Changing practices in faculty evaluation: Can better evaluation make a difference? *Academe, 100*(3), 35–38 May/June 2014.
Miller, M. H. (2010). Online evaluations show same results, lower response rate. *The Chronicle of Higher Education*, May 6, 2010. Retrieved from http://www.chronicle.com/blogs/wiredcampus/online-evaluations-show-same-results-lower-response-rate/23772.
Moriarty, T. A. (2009). They love me, they love me not. *The Chronicle of Higher Education*, April 24, 2009. Retrieved from http://www.chronicle.com/article/They-Love-Me-They-Love-Me-Not/2363.
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.
National Public Radio. (2015). What if students could fire their professors? April 26, 2015. Retrieved from http://www.npr.org/sections/ed/2015/04/26/401953167/what-if-students-could-fire-their-professors.
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 K. Gillespie, & D. Robertson (Eds.), *A guide to faculty development* (pp. 3–20), 2nd ed. San Francisco, California: Jossey-Bass.
Pallett, W. H. (2006). Uses and abuses of student ratings. In P. Seldin (Ed.), *Evaluating faculty performance* (pp. 50–65). Bolton, Massachusetts: Anker.
Patton, S. (2015). Student evaluations: feared, loathed, and not going anywhere. *The Chronicle of Higher Education*, May 19, 2015. Retrieved from http://chroniclevitae.com/news/1011-studentevaluations-feared-loathed-and-not-going-anywhere.
Perlmutter, D. D. (2011). How to read a student evaluation. *The Chronicle of Higher Education*, October 30, 2011. Retrieved from http://www.chronicle.com/article/How-to-Read-a-Student/129553.
Pettit, E. (2016). How one professor is trying to paint a richer portrait of effective teaching. *The Chronicle of Higher Education*, June 16, 2016. Retrieved from http://www.chronicle.com/article/How-One-Professor-Is-Trying-to/236827.
Powers, E. (2007). Sweetening the deal. *Inside Higher Education*, October 18, 2007. Retrieved from https://www.insidehighered.com/news/2007/10/18/sweets.
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.
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.
Ryalls, K., Benton, S., Barr, J., & Li, D. (2016). Response to bias against female instructors. *IDEA Research and Papers*. Editorial Notes.
Schuman, R. (2014). Needs improvement: student evaluations of professors aren't just biased and absurd—they don't even work. *Slate*, April 24 2014. Retrieve from http://www.slate.com/articles/life/education/2014/04/student_evaluations_of_college_professors_are_biased_and_worthless.html.
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 Bulletin, 26*(11), 1329–1342.
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. (2009). Student ratings of teaching effectiveness for faculty groups based on race and gender. *Education, 129*(4), 615–624.

Smith, B. P., & Hawkins, B. (2011). Examining student evaluations of black college faculty: Does race matter? *The Journal of Negro Education*, *80*(2, Spring), 149–162.
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.
Sprague, H. (2016). The bias in student course evaluations. *Inside Higher Education*, June 17, 2016. Retrieved from https://www.insidehighered.com/advice/2016/06/17/removing-bias-student-evaluations-faculty-members-essay.
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. (1990). Editors Notes. *New Directions for Teaching and Learning*, *43*, 1–14.
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. (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.
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.
Warner, J. (2012a). Student evaluations: Part 1 of 2 (probably). *Inside Higher Education*, December 19, 2012. Retrieved from https://www.insidehighered.com/blogs/just-visiting/student-evaluations-part-1-2-probably.
Warner, J. (2012b). Student evaluations: Part 2 of 2. *Inside Higher Education*, December 20, 2012. Retrieved from https://www.insidehighered.com/blogs/just-visiting/student-evaluations-part-2-2.
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.
Weir, R. (2010). Evaluating Evaluations. *Inside Higher Education*, January 4, 2010. Retrieved from https://www.insidehighered.com/advice/2010/01/04/evaluating-evaluations.
Wilson, R. C. (1986). Improving faculty teaching: Effective use of student evaluations and consultants. *The Journal of Higher Education*, *57*(2), 196–211.
Zaino, J. (2015). Gender bias in student evaluations. *Inside Higher Education*, February 23, 2015. Retrieved from https://www.insidehighered.com/blogs/university-venus/gender-bias-student-evaluations.
Zakrajsek, T. D. (2010). Important skills and knowledge, In K. Gillespie, & D. Robertson (Eds.), *A guide to faculty development* (pp. 83–98), 2nd ed. San Francisco, California: Jossey-Bass.
Zubizarreta, J. (1999). Evaluating teaching through portfolios. In P. Seldin (Ed.), *Changing practices in evaluating teaching* (pp. 162–182). Bolton, Massachusetts: Anker.