Reducing Bias and Improving Benefit in Evaluation of Teaching

Dr. Naomi C. Chesler, University of Wisconsin, Madison

Naomi C. Chesler is Professor of Biomedical Engineering with an affiliate appointment in Educational Psychology. Her research interests include vascular biomechanics, hemodynamics and cardiac function as well as the factors that motivate students to pursue and persist in engineering careers, with a focus on women and under-represented minorities. More information is available at the website for the Vascular Tissue Biomechanics lab at UW-Madison: vtb.engr.wisc.edu

Dr. Dante Fratta, University of Wisconsin, Madison
Elizabeth C Harris, University of Wisconsin, Madison

Elizabeth Harris has been part of the University of Wisconsin Madison's College of Engineering since 2012. She approaches Engineering Education opportunities by leveraging her background in cognitive and systems engineering in addition to her background in education. Her work focuses on improving the effectiveness of the Institution, and the experiences of students, faculty, and academic staff, by addressing the holistic ecologies present around teaching and learning at UW Madison. She does this by partnering to foster, create, and explore cultural and strategic interventions, in addition to practical.

Prof. Wayne P. Pferdehirt, University of Wisconsin, Madison

Wayne P. Pferdehirt has directed several online graduate engineering degree programs for practicing engineers within the University of Wisconsin-Madison's College of Engineering since 1998. Wayne serves as a member of the College's Education Innovation Committee and chairs the College's Master of Engineering Oversight Committee. Wayne is a frequent speaker and author on continuing education for engineers, project management, and engineering leadership.

Heidi-Lynn Ploeg, Queens University at Kingston

Dr. Ploeg is an Associate Professor of Mechanical and Materials Engineering at Queen's University at Kingston, Ontario. From July 2003 to August 2018 she was an Assistant and Associate Professor in Mechanical Engineering with affiliate appointments in Biomedical Engineering, Material Science & Engineering, and Orthopedics & Rehabilitation, at the University of Wisconsin-Madison, where she established and directed the Bone and Joint Biomechanics (BJB) Laboratory. Dr. Ploeg received her Ph.D. in Mechanical Engineering from Queen's University at Kingston, Ontario, Canada in 2000. She was the Director Preclinical Stress Analysis Group in the Research Department at Sulzer Orthopedics Ltd. (now Zimmer-Biomet GmbH), Winterthur, Switzerland from 1992-2002. Dr. Ploeg's research focus is orthopedic biomechanics including design of medical devices, bone modeling and remodeling, mechanical testing, and finite element modeling.

Prof. Barry D. Van Veen, University of Wisconsin, Madison

Barry Van Veen received the B.S. in Electrical Engineering from Michigan Technological University and the Ph.D. in Electrical Engineering from the University of Colorado, Boulder. He currently is the Lynn H. Matthias Professor and Associate Chair for Graduate and Online Studies in the Electrical and Computer Engineering Department at the University of Wisconsin-Madison. He has received multiple teaching awards for development and implementation of active learning methods in signal processing and machine learning classes.

Reducing bias and improving benefit in evaluation of teaching

Abstract

Engineering teaching assessment at the college-level should provide: 1) data to assess the quality of instruction provided by an instructor; 2) instructors with actionable information on how their instruction may be improved; and 3) evidence of effective instruction for tenure and promotion purposes. Many institutions rely primarily on student evaluations of teaching (SET) for teaching assessment. Peer evaluations of teaching are rarely used outside of the tenure evaluation period for assistant professors. Recent research has provided compelling evidence that SET have significant systemic bias with respect to gender, race, and sexual orientation and moreover do not assess teaching effectiveness. These biases and limitations indicate that SET should be used with caution in promotion processes and to revise instructional practices. Peer evaluations as typically implemented are of limited utility in evaluating or improving instruction, due in part to limited awareness of best practices by evaluators and in part due to the time investment required for a thorough evaluation. Here, we report on the biases and problems with SET and present a guide for peer evaluation that could improve formative feedback and assessment of engineering instruction.

Introduction

Many Schools and Colleges of Engineering collect student evaluations of teaching (SET) for courses and then use the compiled data to evaluate the performance of faculty, staff and student instructors. However, recent studies call this use of SET into question. In particular, evidence has shown that SET do not reflect the effectiveness of instruction or learning [1], [2]. Therefore, changes to teaching practices in response to student feedback through SET may not improve teaching or learning effectiveness. Moreover, SET data are biased. Multiple studies have shown that SET results are biased with respect to gender [3], to sexual orientation and gender identity [4], and to race [5]. Even purportedly objective metrics, such as how quickly homework are graded and returned, are affected by student understanding of gender [3]. For all of these reasons, SET are problematic when used in merit, tenure and promotion decisions.

Appropriate Use of SET

SET are unlikely to be eliminated since they are an efficient way to gather student feedback. Also, despite their limitations, SET can yield useful information. For example, very poor evaluations can be indications of problems that must be addressed by the instructor or the institution [6]. In addition, SET can be used by individual instructors teaching the same course in subsequent years to track student response to instructional changes. Importantly, SET provide student feedback. Rather than using these data to evaluate instruction, SET can be a useful tool to gather feedback on students' expectations [3]. The individual student comments from SET can guide course and instructor improvement. Stark and Freishtat [1] provide a set of recommendations on appropriate use of SET:

- effectiveness and value of courses should not be used as evaluation items,
- averages of scores should be de-emphasized while score distribution and number of responses should be reported,
- results from a low number of responses should not be used,
- comments from students should be considered while understanding their limitations, and

- SET comparisons between different courses' content, types, sizes and areas of studies should be avoided - e.g., departments should not compare individual instructors against departmental averages

Regarding averages of scores in particular, disadvantages as compared to distributions include that:

- They presume the differences between values at the lower end of the scale (e.g., a 1 and a 2) is the same as that between values at the higher end of the scale (e.g., a 4 and a 5);
- They presume that the difference between numerical ratings (e.g., 3 and 4) mean the same to different students;
- They presume that a given numerical rating (e.g., 4) means the same thing to different students;
- They presume a rating of 5 balances a 3 to be equivalent to two ratings of 4; and
- At least half the faculty in any department will have average scores below the median average score.

In summary, the use of averages promotes simplistic judgement of a complex, nuanced activity.

An illustration of recommended use of SET is provided in Appendix 1. The example compares the instructor to the department using distributions of scores, and response rates are given. Written student comments can be useful and informative, but must be used with caution since they often reflect implicit bias. Finally, as noted above, while SET data can be useful to individual instructors teaching the same course in subsequent years to assess changes in student feedback over time, great caution must be exercised when using SET to compare the teaching effectiveness of different faculty members teaching the same course.

Alternatives to SET for Student Feedback

An alternative form of student feedback on teaching effectiveness can be obtained using focus groups [7]. A focus group involves a neutral party facilitating a conversation about the instructor's teaching with a group of current or former students. It is important that the facilitator be trained in conducting focus groups so that the student conversation is productive and provides useful specific information (see e.g., Facilitator Tool Kit http://oqi.wisc.edu/resourcelibrary/uploads/resources/Facilitator_Tool_Kit.pdf).

Retrospective evaluations are similar to focus groups and occur at some time after the end of the course. However, these forms of collecting student feedback are time- and resource-intensive. Few institutions have sufficient personnel trained to properly conduct this type of qualitative data collection at the scale necessary.

Peer Evaluation for Teaching Assessment

One frequently used tool for assessing teaching effectiveness is observation by peers [8]. In Peer Evaluation of Teaching, a peer of the instructor observes one or several classes and provides feedback regarding learning and teaching activities. The systematic collection of peer evaluations of teaching are then documented in dossiers used for promotion and tenure decisions. One of the limitations of peer evaluation is that it takes a significant effort to obtain useful information. For example, Weimer [9] summarizes a set of recommended practices needed for effective peer evaluation from Chism [10]. These include:

- Peers need to be trained for the effective use of the technique.
- A single class observation is not enough to obtain reliable indicators of teaching/learning effectiveness. At least three observations are recommended.
- Contextual information regarding the classroom activity, such as learning objectives, must be provided to the observer.
- A checklist must be provided to help the peer assess specific areas of the learning/teaching activity.
- The peer should be a strict observer. Participation of the peer in the teaching/learning activity – even their presence in the room – may distract the instructor or students from their activity.
- The peer should observe entire class periods.
- A written review should be provided right after the observed session.

Self-Assessment of Teaching

One aspect of continuous improvement that is not typically included in peer evaluations of teaching is a self-assessment. Reflection is key to learning in many professional domains, including teaching [11]. As instructors become more experienced, self-assessment and self-reflection – for example through journaling, watching videos of themselves in the classroom or other evaluations – become useful forms of ongoing professional development [12]. To guide self-assessment and reflection in mathematics and science instruction, Wieman and Gilbert developed a Teaching Practices Inventory [13]. The authors argue that teaching effectiveness can be measured by characterizing the use of best practices. In developing this inventory, Weiman and Gilbert sought to address the following objectives: validity–the result of the assessment must be strongly correlated with achievement; meaningful comparisons–the instructor should be able to be compare their performance against peers; fairness–the tool should be widely applicable and valid across curricula; practicality–implementation should be inexpensive for instructors and institutions; and improvement–the results could be used by instructor to improve their teaching. Currently, the TPI is available in multiple formats: paper, Excel file, and anonymous Qualtrics survey (http://www.cwsei.ubc.ca/resources/TeachingPracticesInventory.htm) and takes ~10 min to complete. Unlike a grading system in which 100 is the best and 0 is the worst, scores of 100% would indicate that all best-practices are used in a single course, which is not encouraged by the authors. In a single high-performing department studied by Wieman and Gilbert, the distribution of scores ranged from 10 to 50 with a median in the low 30s [6]. Indeed, the numerical score is not the emphasis here; rather the value of this inventory is that the instructor reflects on their current practice and considers how best-practices may be incorporated to improve teaching.

A Proposal for Improved Peer Evaluation

Building on the TPI for self-assessment and the known limitations of peer evaluation of teaching, we recommend the following series of steps for improved peer evaluation of teaching:

1) The instructor is provided with resources for teaching development, self-assessment, and the Peer Evaluation Guide (Appendix 2) at least three months prior to the beginning of the semester so that these materials drive course design and/or revision.
2) The instructor completes a self-assessment and writes a brief, reflective narrative explaining their rationale for the practices implemented in their course.
3) The evaluator and instructor schedule at least one classroom observation.
4) The instructor provides the evaluator with the narrative and access to their course materials at least two weeks prior to the classroom observation. The instructor may request a meeting with the evaluator to provide additional explanation prior to any classroom observation.
5) The evaluator assesses the provided materials and classroom instruction using the Peer Evaluation Guide.
6) The evaluator meets with the instructor to provide informal feedback, taking the opportunity to ask and answer questions.
7) The evaluator writes a formal assessment of the instructor's teaching practices and submits the assessment within three weeks of the classroom observation.

Summary

Teaching and learning are complex activities. Thus, it is no surprise that they are difficult and time-consuming to assess properly. When gathering student feedback, SET are an alluring tool: there is an automated process for obtaining them and they can be improperly used to distill assessment of teaching for an instructor of a given course to a single number. However, significant gender, race, and sexual orientation bias in SET and evidence that SET do not measure learning raise serious concerns about continued reliance on them for merit and promotion decisions. MacNell et al. [3] among other researchers, suggest that universities consider phasing out SET for tenure and promotion decisions "but still use them to get feedback on what students want and expect from their courses."

Use of research-proven best practices for teaching should also be considered for promotion and tenure decisions given their correlation with teaching effectiveness [13].

We recommend that departments encourage their instructors, particularly probationary faculty, to take advantage of the wide range of institutional resources for improving teaching. Appropriate use of research-proven best practices will improve the quality of teaching and learning. It will also increase the success and job satisfaction of instructors. The process of improved peer evaluation of teaching proposed here addresses key limitations of the most current peer evaluation processes and may lead to more inclusion of best practices of teaching by instructors of engineering.

References

[1] Stark, P. B. and Freishtat, R. (2014), An Evaluation of Course Evaluations, Science Open Research, DOI: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1.

[2] Boring, A., Ottoboni, K., and Stark, P. B. (2016), Student evaluations of teaching (mostly) do not measure teaching effectiveness, Science Open Research, DOI: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1.

[3] MacNell, L., Driscoll, A., and Hunt, A., N. (2015), What’s in a Name: Exposing Gender Bias in Student Ratings of Teaching, Innov High Educ, 40: 291-303, DOI: 10.1007/s10755-014-9313-4.

[4] Ewing, V. L., Stukas, Jr., A. A., and Sheehan, E. P., (2003). Student Prejudice Against Gay Male and Lesbian Lecturers. The Journal of Social Psychology, 143(5): 569-579.

[5] Ho, A. K., Thomsen, L., and Sidanius, J. (2009), “Perceived Academic Competence and Overall Job Evaluations: Students’ Evaluations of African American and European American Professors.” J of App Soc Psych 39(2): 389-406.

[6] Wieman, C. (2015), A Better Way to Evaluate Undergraduate Teaching. Change, The Magazine of Higher Learning, January-February 2015.

[7] Simon, J.S. (1999), The Wilder Nonprofit Guide to Conducting Successful Focus Groups. Amherst H. Wilder Foundation.

[8] Hill, H. C. and Grossman, P. (2015), Learning from Teacher Observations: Challenges and Opportunities Posed by New Teacher Evaluation Systems, Harvard Educational Review. 83(2): 371-384.

[9] Weimer, M. (2010), Guidelines for Effective Classroom Observations, URL: http://www.facultyfocus.com/articles/faculty-evaluation/guidelines-for-effective-classroom-observations/ [Accessed on March 23, 2017].

[10] Chism, N.V.N. (2007). *Peer Review of Teaching: A Sourcebook.* 2nd Ed. Bolton, Mass.: Anker.

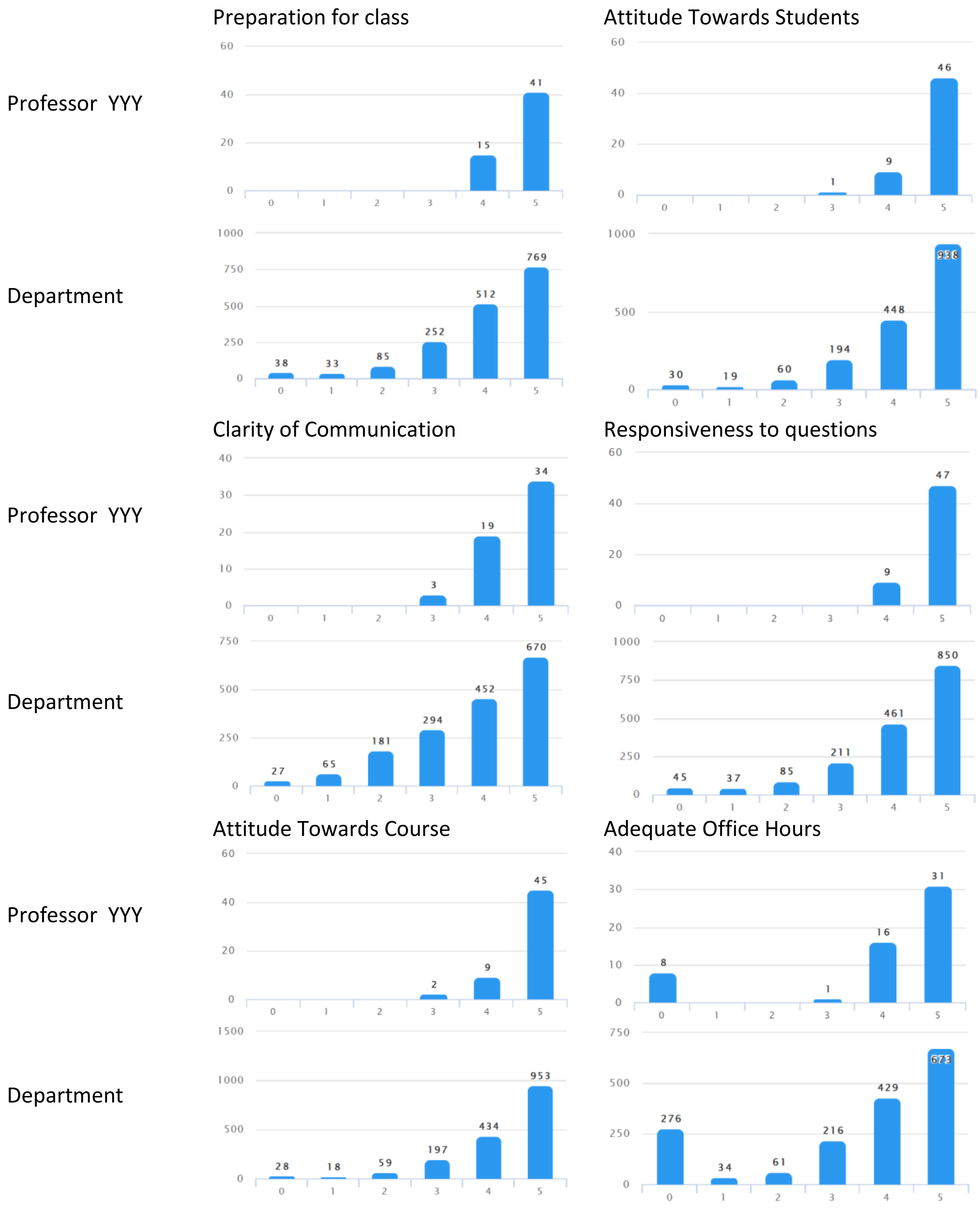
[11] Schön, (1983) *The Reflective Practitioner: How professionals think in action*, Temple Smith

[12] Zeichner and Liston (2013), *Reflective teaching: An introduction*, Routledge

[13] Wieman, C. and Gilbert, S. (2014), The Teaching Practices Inventory: A New Tool for Characterizing College and University Teaching in Mathematics and Science, CBE—Life Sciences Education, 13: 552–569.

Appendix 1. Sample Presentation of Student Evaluations of Teaching

Professor YYY for ECEXXX. 81 Students, 56 Responses; Response rate 69%. (Dept response rate 52%)



Preparation for class

Attitude Towards Students

Clarity of Communication

Responsiveness to questions

Attitude Towards Course

Adequate Office Hours

Appendix 2. Guide for Peer Evaluation

The evaluator's letter should assess appropriate use of the evidence-based best teaching practices described in Tables A and B based on: review of course materials; classroom observation; and, discussion with the instructor. The practices in Table A have been shown to have a particularly large and robust impact on learning in undergraduate STEM courses (see, e.g., Wieman and Gilbert, 2014). The literature on teaching indicates that practices in Table B promote learning also, but with less impact than those in Table A. Hence, evaluators should pay particular attention to the practices in Table A that the instructor chooses to implement in their course.

Not all practices described in Tables A and B are appropriate for every course. Wieman and Gilbert identified these as best practices for undergraduate non-laboratory courses. However, laboratory and graduate courses may benefit from many of the practices listed. The instructor is expected to use their professional judgement in choosing to implement practices that will result in the best learning outcomes for the students. All practices they choose to employ should be implemented thoughtfully. Furthermore, instructors are expected to show development in teaching over time as evidenced by increased breadth and depth of skill.

Table A. Practices with Large and Robust Impact on Learning

Practice	Assessment criteria
A.1. Provide specific and measurable learning objectives (or outcomes)	- Learning objectives are provided to the students - Learning objectives are clearly written from the student's perspective to describe what the student will be able to accomplish upon completion of the course - The learning objectives are specific and measurable - The learning objectives are consistent with the course catalog description
A.2. Use active learning	- The instructor is typically lecturing (presenting content, deriving results, showing problem solutions, etc.) less than sixty percent of the class time - Students are engaged in small group discussions, problem solving, or design exercises during class - Students are asked to read/view material on an upcoming class session and complete assignments or quizzes on it shortly before or at the start of class
A.3. Provide formative feedback	- Problem sets/homework are counted toward course grade and assigned at intervals of two weeks or less - Encouragement and facilitation for students to work collaboratively on their assignments - Feedback is provided before grading or with opportunity to redo work to improve grade
A.4. Provide summative feedback	- A minimum of two midterm exams or assessments
A.5. Other	- An instructor-independent pre/post-test (e.g., concept inventory) is used to measure learning - A consistent measure of learning is repeated in multiple offerings of the course and used to compare learning - New teaching methods or materials are tried along with measurements to determine their impact on student learning - Instructor and TA(s) or coaches meet at least every two weeks to discuss student learning, challenges, and upcoming material

Table B. Practices that Support Learning

Category	Practice
B.1. Materials provided to students	- List of competencies that are not topic-related (critical thinking, problem solving…) are described - Affective goals are provided to influence attitudes and beliefs (interest, motivation, relevance, beliefs about competencies, how to master material) - Solutions to homework problems, practice or previous years' exams, worked examples, lecture notes are provided - Articles from scientific literature are used in the class
B.2. In-class features	- Use of one-minute paper or other reflective activity is used in class - The average number of classes where demonstrations, simulations or video are shown AND students first record predicted behavior and afterwards explicitly compare observations with predictions is greater than one-half - Students give presentations (oral or poster)
B.3. Assignments	- A paper or project taking longer than two weeks and involving some degree of student control in choice of topic or design is assigned - Explicit group assignments are given
B.4. Feedback from students to instructor	- Students are asked to provide one or more midterm course evaluation - Repeated online or paper feedback or via some other means such as clickers is used
B.5. Feedback to students	- Students see marked assignments - Students see assignment answer key and/or marking rubric - Students see marked midterm exams - Students see midterm exam answer keys - Students are explicitly encouraged to meet with you - Students receive feedback in a timely manner
B.6. Testing and grading	- Greater than 15% of exam score is assigned to questions that require students to explain reasoning - Less than 60% of the course grade is assigned to the final exam
B.7. Other	- An assessment is given at beginning of course to measure background knowledge - A pre-post survey of student interest and/or perceptions about the subject is used - Opportunities are provided for students' self-evaluation of learning - Students are provided with opportunities to have some control over their learning such as choice of topics for course, paper, or project, choice of assessment methods, etc. - The instructor discussed how to teach the course with colleagues three or more times - The instructor read literature about teaching and learning relevant to this course - The instructor sat in on a colleague's class (any class) to get/share ideas for teaching two or more times