ASEE 2022 ANNUAL CONFERENCE
Excellence Through Diversity
MINNEAPOLIS, MINNESOTA, JUNE 26TH-29TH, 2022 ASEE

Paper ID #37195

Work in Progress: Updating End of Semester Course Evaluations via Backwards Design to Reduce Student Bias

Adam St. Jean (Associate Teaching Professor of Biomedical Engineering)

Adam St. Jean is an Associate Teaching Professor and the Associate Chair for Undergraduate Programs in Biomedical Engineering at UMass Lowell. He received his Ph.D. in Chemical Engineering from the University of Massachusetts Amherst in 2012. His current research interests include scientific literacy and engineering identity/experiences for the LBGTQ+ community.

Yanfen Li (Assistant Teaching Professor)

Yanfen Li is an Assistant Teaching Professor at the University of Massachusetts Lowell. She received her PhD in Bioengineering from the University of Illinois at Urbana Champaign. Her current research is in engineering education with a focus on curriculum development and retention of female and minority students in engineering.

Chiara Ghezzi

Assistant Professor and Associate Chair

Laura Punnett

Work in Progress: Updating End of Semester Course Evaluations via Backwards Design to Reduce Student Bias

Abstract

Many universities conduct course evaluations at the end of the semester to evaluate the quality of teaching from an instructor. These evaluations are often used for consideration of tenure, compensation, employment decisions, and teaching awards, among other career milestones. However, a variety of literature indicates that student evaluations of teaching may not be an accurate indication of teaching effectiveness [1], [2]. In particular, student biases about factors such as gender, race, and age can all affect their evaluations [3]–[5]. In this Work-in-progress article, we introduce a backwards design approach to re-evaluate the use and goals of course evaluations from multiple stakeholders including faculty, administrators, and students. These goals are then used to redefine the types of questions needed in course evaluation questionnaires. We also introduce a new method of writing questionnaire questions to be evidence-based (i.e., did the instructor grade assignments within a week) rather than intuition-based (i.e., did the instructor grade assignments in a timely manner) to reduce student bias.

Background

End-of-semester course evaluations are a commonplace tool used by faculty and universities to solicit student feedback on instructor and course quality. Since students are the direct targets of faculty services, there is an authentic need to include their comments in any assessment of teaching. This feedback is used in myriad important ways, including formative assessment to improve instructor skill and summative assessment for promotion and tenure processes, awards, compensation, and employment decisions. Given the weight that this feedback can play in very important processes, it is essential that the collected feedback be as accurate and unbiased as possible.

Students, faculty, and administrators all benefit from accurate measurement of teaching effectiveness but it is difficult to achieve. There is a variety of literature indicating that Student Evaluations of Teaching (SET) may not be an accurate indicator of teaching effectiveness [1], [2] due to several factors. Course evaluation instruments may contain overly broad or subjective language that is assumed to capture the desired data or may contain academic jargon that is confusing to students. Student responses to such questions can vary widely due to a lack of student calibration or calibration inconsistencies. Additionally, student biases towards factors such as gender, race/ethnicity, and age can affect their evaluations [3]–[5]. These biases could then negatively impact faculty's career trajectory, tenure, and promotion.

The belief is widespread among faculty members that student evaluations are not accurate. According to posts on online platforms such as blogs, YouTube, etc., this can lead to feedback being disregarded by faculty; this negates the formative purpose of the assessment. Further, faculty motivation to improve teaching may be hindered if they know or believe that their efforts may not be reflected by the SET. Worse, negative reviews that are inaccurate could negatively impact faculty confidence or mental health, potentially creating barriers to career success [6]. Promisingly, some work has shown that simple interventions, such as including pre-survey language to make students aware of implicit biases, can begin to mitigate bias in SET [7].

In this project, we propose to use a backwards design approach [8], [9] to redesign the SET instruments used in the Biomedical Engineering (BME) Department at the University of Massachusetts Lowell (UML) to determine whether actionable changes in SET language can subtly reduce effect of student bias.

Re-evaluation and Design Process

Backwards design is a process used in education to create learning experiences that accomplish specific learning objectives. In brief, this approach requires deliberate consideration of desired learning objectives prior to the development of assessments and meaningful learning experiences instead of alternatively designing lessons based on content without consideration of the learning goals. This methodology has been used to inform other design processes separate from education [10]. A group of tenure-track, tenured, teaching faculty and chairs aimed to use this framework to analyze the goals of course evaluations from multiple stakeholders. These goals are then used to redefine the type and style questions needed in course evaluation questionnaires. Specifically, we introduce a new method of writing SET questions to be evidence-based (objective) rather than intuition-based (subjective) to reduce student biases.

First, we examined the distinct set of goals that each of the main stakeholders would have for using or providing the information collected by SET. For faculty, the main goals include course improvement, pedagogical improvement, promotion/tenure, merit pay, and awards. Administrators primarily use SET data for faculty evaluation and promotion/tenure but could also use the information for resource allocation [11]. Students' goals include a sense of voice in their education and improved learning through development of teaching and course design. These goals were then confirmed via preliminary interviews with representative stakeholders. Given these different goals, our team examined what essential questions we wanted to probe via SET (Table 1) and subsequently solicited input from stakeholder representatives. These questions are not those included in the questionnaires but are rather the guiding questions for the backwards design process.

Table 1 SET Content Areas

Content Area Essential Questions		Stakeholders
A	To what extent did the students learn the content contained in the learning objectives?	Faculty, Student
B	To what extent did the course meet ABET student outcomes?	Faculty, Administrator
C	Was the way(s) the course was taught effective at helping students learn the content we intended for them to learn?	Faculty, Administrator, Student
D	To what extent did instructor put effort into making their course effective?	Faculty, Administrator
E	To what extent was the instructor innovative in their teaching?	Faculty, Administrator
F	Did the students have the proper background to take this course?	Faculty, Student
G	Were class resources/facilities adequate for the course?	Faculty, Administrator, Student
H	To what extent is does the instructor create a quality and inclusive learning environment?	Faculty, Administrator, Student

Evaluation Results for Current SET

To evaluate the state of our current system, we reviewed the current SET questions in use in the BME department at UML and coded them based on which essential question they addressed (Table 2). As part of this review, we also identified broad and subjective language in the questions which were likely to allow or possibly encourage bias. The SET questions only addressed a very small subset of the essential questions we were trying to probe. Nearly all of them covered the students' perception of the effort that faculty were putting into courses, and none reflected an assessment of innovative teaching, student learning, or other considerations. The open-response questions did allow the students to reflect more broadly on the instructor and the course. Based on the experiences of the authors, these sections tend to contain students' comments on aspects of the course that they found useful in their learning. However, the SET was generally found to be lacking in appropriate questions to probe these essential questions in a meaningful way. Overall, the limited scope of the questions and inclusion of subjective language reduces the authenticity of the instrument as a useful tool for assessment.

In addition, 14 of the 17 rating-based questions contained language that could potentially be subject to biased rating. Students are not calibrated on the relative meaning of the words and are therefore not able to objectively identify if an instructor was 'adequate,' 'timely,' or 'fair.' For example, specific course attributes (e.g., size, course type, teaching support) often affect how quickly assignments can be graded. When uncalibrated students implicitly compare faculty across different courses, those faculty with smaller courses might receive higher ratings on the "timely grading" question (Q12) simply because their course was smaller. Similarly, if one instructor is giving deep, meaningful feedback on open-ended assignments and another is using computer-graded assessments with little feedback, the student may give higher timeliness ratings to the latter. Ultimately, the system may not reflect or reward faculty who are innovating or providing excellent teaching in favor of those who are able to obtain higher scores regardless of the underlying reasons.

Implicit student biases may also influence their assessment of a particular faculty member based on gender, race, age, or other attributes [1], [2], [5], [12]–[15]. Recent work shows that these effects are confounded [16], potentially creating greater disparity . To reduce or eliminate these effects, using clear, objective language is preferable. For example, in Q13, "Grading of exams and assignments was done in a fair manner" could be rephrased to "Exams and assignments were graded using provided rubrics." Removing the phrases requiring interpretation facilitates a clearer, more objective assessment.

Table 2 Current list of SET questions. Unless otherwise specified, all question responses are submitted using a 5-point Likert scale (Strongly Disagree-Strongly Agree). Red color identifies language that may be subjective.

Essential Question Code	**SET questions**
D	1. The course instructor was well organized.
D	2. The course instructor displayed adequate knowledge of subject matter.
D	3. The course instructor was punctual in starting and ending the class.
D	4. The course instructor provided clear explanations of the course material.
D	5. The course instructor behaved in a professional manner.
D	6. The course instructor encouraged interaction and questions from the students.
D	7. The course instructor answered questions satisfactorily.
D	8. The course instructor used the board/visuals/computer effectively.
D	9. The course instructor spoke loudly and clearly.
D	10. The course instructor established and maintained office hours.
D	11. The course instructor distributed a course syllabus at the start of the semester that specified the course expectations, grading criteria, and student responsibilities.
D	12. The course instructor corrected assignments and exams in a timely manner.
D	13. Grading of exams and assignments was done in a fair manner.
D	14. Course workload was appropriate.
D	15. Course instructor had good command of the class and control over disruptive student behavior.
D	16. Course Instructor treated students with respect.
D	17. Overall instructor rating.
E	18. What aspects of the course was most useful? **(open response):**
E	19. What would you suggest to improve this course (content, structure, grading, etc.)? **(open response):**
E/G	20. What would you suggest to the instructor to improve the teaching of this course? **(open response):**
E	21. Any other comments? **(open response):**

Next Steps

Given the assessment of our current SET instrument, there is much opportunity to reduce the influence of implicit biases and expand its usefulness to all stakeholders. There are two main goals for the next phase of the project: 1) create/revise questions to align with all essential questions, and 2) eliminate subjective and include objective language in each question.

For each of these goals, we are in the process of determining the best assessment methodology. We intent to collect feedback at multiple phases from all stakeholders to assess clarity, objectivity, and alignment with essential questions. Once complete, we will pilot the new instrument and evaluate its effectiveness in reducing bias and more appropriately assessing effective teaching and learning.

Citations

[1] A. Boring, K. Ottoboni, and P. Stark, "Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness," *Sci. Res.*, pp. 1–11, 2016, doi: 10.14293/s2199-1006.1.sor-edu.aetbzc.v1.

[2] B. Uttl, C. A. White, and D. W. Gonzalez, "Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related," *Stud. Educ. Eval.*, vol. 54, pp. 22–42, 2017, doi: 10.1016/j.stueduc.2016.08.007.

[3] J. Arbuckle and B. D. Williams, "Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations," *Sex Roles*, vol. 49, no. 9–10, pp. 507–516, Nov. 2003, doi: 10.1023/A:1025832707002.

[4] Y. Fan *et al.*, "SAE Notes: Student Satisfaction Statement," *PLoS One*, vol. 14, no. 2, pp. 1–16, 2019.

[5] D. Storage, Z. Horne, A. Cimpian, and S. J. Leslie, "The frequency of 'brilliant' and 'genius' in teaching evaluations predicts the representation of women and African Americans across fields," *PLoS One*, vol. 11, no. 3, pp. 1–17, 2016, doi: 10.1371/journal.pone.0150194.

[6] R. Lakeman *et al.*, "Appearance, insults, allegations, blame and threats: an analysis of anonymous non-constructive student evaluation of teaching in Australia," *https://doi.org/10.1080/02602938.2021.2012643*, 2021, doi: 10.1080/02602938.2021.2012643.

[7] D. A. M. Peterson, L. A. Biederman, D. Andersen, T. M. Ditonto, and K. Roe, "Mitigating gender bias in student evaluations of teaching," *PLoS One*, vol. 14, no. 5, pp. 1–11, 2019, doi: 10.1371/journal.pone.0216241.

[8] G. Wiggins, G. P. Wiggins, and J. McTighe, *Understanding by Design*, 2nd ed. Association for Supervision and Curriculum Development, 2005.

[9] J. McTighe and R. S. Thomas, "Backward Design by Forward Action," *Educ. Leadersh.*, vol. 60, pp. 52–55, 2003.

[10] B. Kantorski, C. W. Sanford-Dolly, D. R. Commisso, and J. A. Pollock, "Backward design as a mobile application development strategy," *Educ. Technol. Res. Dev.*, vol. 67, no. 3, pp. 711–731, Jun. 2019, doi: 10.1007/S11423-019-09662-7/FIGURES/8.

[11] A. R. Linse, “Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees,” *Stud. Educ. Eval.*, vol. 54, pp. 94–106, Sep. 2017, doi: 10.1016/J.STUEDUC.2016.12.004.

[12] J. Esarey and N. Valdes, “Unbiased, reliable, and valid student evaluations can still be unfair,” *Assess. Eval. High. Educ.*, vol. 45, no. 8, pp. 1106–1120, 2020, doi: 10.1080/02602938.2020.1724875.

[13] Y. Fanid *et al.*, “Gender and cultural bias in student evaluations: Why representation matters,” 2019, doi: 10.1371/journal.pone.0209749.

[14] D. A. M. Peterson, L. A. Biederman, D. Andersen, T. M. Ditonto, and K. Roe, “Mitigating gender bias in student evaluations of teaching,” *PLoS One*, vol. 14, no. 5, pp. 1–10, 2019, doi: 10.1371/journal.pone.0216241.

[15] N. Punyanunt-Carter and S. L. Carter, “Students’ Gender Bias in Teaching Evaluations,” *High. Learn. Res. Commun.*, vol. 5, no. 3, p. 28, 2015, doi: 10.18870/hlrc.v5i3.234.

[16] N. Radchenko, “Student evaluations of teaching: unidimensionality, subjectivity, and biases,” *https://doi.org/10.1080/09645292.2020.1814997*, vol. 28, no. 6, pp. 549–566, Nov. 2020, doi: 10.1080/09645292.2020.1814997.