This material is based upon work supported by the National Science Foundation under Award No. 2303732
https://bradley.edu/ADVANCEBU/

ADVANCE BU Recommendations for Revising Bradley's Evaluation of Teaching

> "Bradley University strives to maintain an academic environment that supports superior teaching, the primary mission of the University" (*Faculty Handbook* 2024: 44).

With more than 1000 extant studies of Student Evaluations of Teaching (SETs), they are one of the most widely studied aspects of higher education (Berk 2018). A meta-analysis of this literature finds problems with the validity and reliability of the instruments, and with the interpretation and use of the results they generate (Arend 2018). Specifically, many studies find no correlation (or even an inverse correlation) between SETs and student learning (Braga, Paccagnella & Pellizzari, 2014; Langbein, 2008). And many scholars, and even publishers of scientifically validated SETs, stress that they are intended for the purposes of formative evaluation (ongoing improvement) rather than summative evaluation (an annual evaluation "score"), and should constitute no more than 30-50% of an overall evaluation of teaching (Arreola, 2000; Benton & Ryalls, 2016; Berk, 2005; Hoyt & Pallett 2018). Moreover, ample research has demonstrated that standardized SETs result in bias against women, people of color, and other marginalized groups (Chávez & Mitchell, 2020; Boring, 2017; Hornstein, 2017; Boring, et al., 2016; Stark & Freishtat, 2014). Not only does this potentially disadvantage such groups in tenure and promotion decisions, but it potentially creates a legal liability for the university under civil rights laws that protect against "arbitrary and capricious tests that discriminate either directly or indirectly (statistically) against members of protected groups" (Wines and Lau 2006: 169).

In sum, as Weaver, et. al (2020) conclude, "Student surveys of courses are at best unreliable or at worst discriminatory methods to evaluate the quality of teaching." The key limitations of SETs identified in the scholarly literature are summarized below.

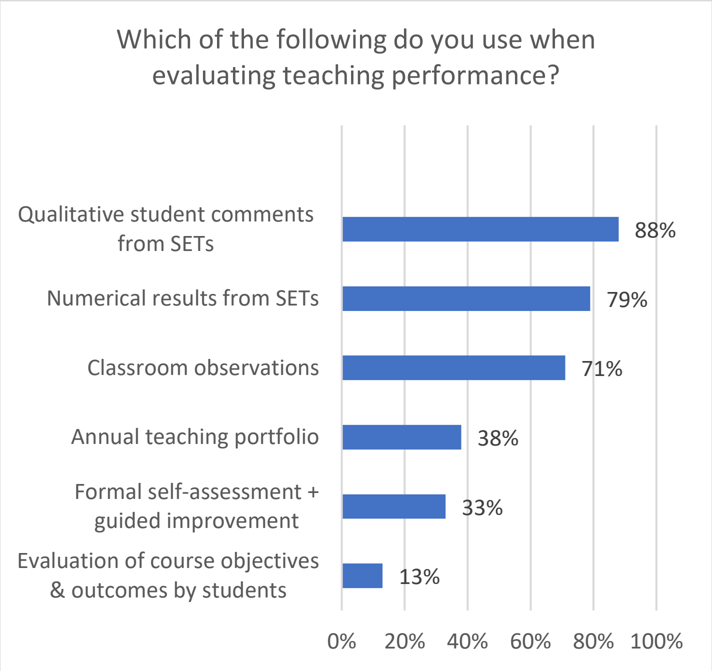
Four Key Limitations of SETs (adapted from Arend 2018)

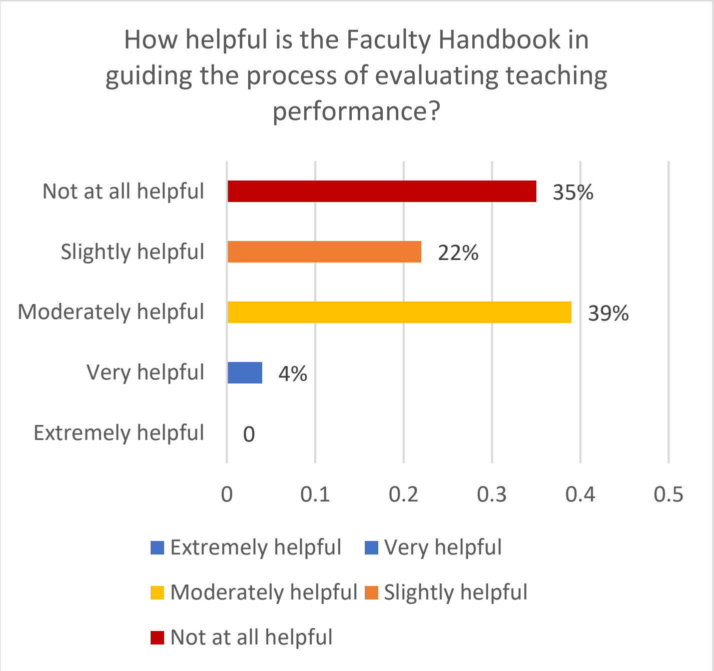
Asking students to comment on matters outside their expertise	Student feedback is crucial. However, most students are not experts in pedagogy or the subject matter of the course. Therefore, they should only be asked questions about which they have expertise, namely their own experiences. (Consistent with this, SETs should be renamed Student Experience Questionnaires or similar.)
Technical inadequacy	(1) Many studies find no positive correlation between SET scores and student learning, and most universities (including Bradley) use home-grown instruments that have not been validated and tested for reliability. (2) Many institutions (including Bradley) do not administer instruments in standardized ways that glean high enough response rates for validity. Low response rates often capture only the most extreme views (the "haters" and the "mega fans"), resulting in bimodal distributions that are difficult to interpret. (3) Some universities (including Bradley) do not use a common set of questions, making comparisons across departments problematic.

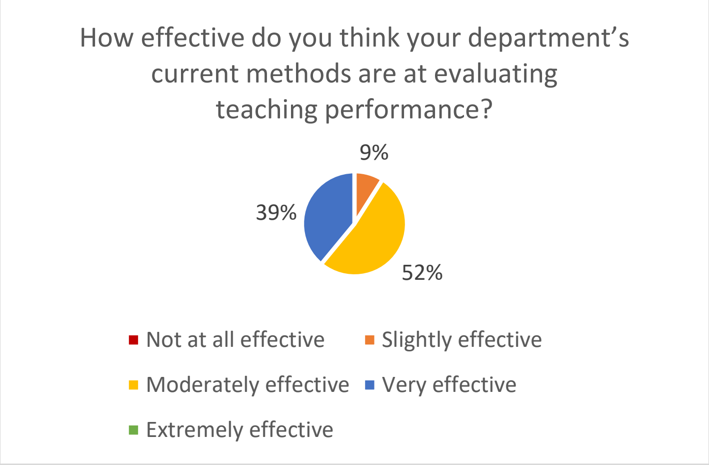
Misinterpretation and misuse of data	Although SET numbers provide a semblance of objectivity and comparability, variations between courses (e.g., small or large classes, introductory or advanced classes, popular or unpopular subjects) mean that comparing the SET numbers for different faculty or different courses is seldom comparing “apples with apples.” While SETs can be a valuable source of formative feedback, allowing the faculty member and chair to see areas for improvement, they are only one piece of evidence of teaching efficacy, and should not be used as the sole evidence for summative evaluation (annual evaluation).
Biased results	Responses on SETs reflect systemic bias against women, faculty of color, and other marginalized groups. More general questions (eg. “Overall, what were the weaknesses of this professor?”) are more subject to bias than questions based on observable behaviors. Marginalized groups are also more likely to receive abusive or discriminatory qualitative comments. Therefore, some universities have systems of review to allow for the removal of such biased responses from the instructor’s record.

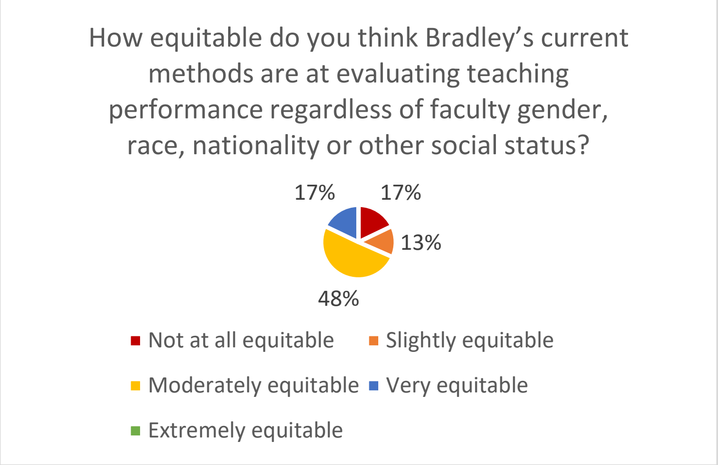
Current Bradley Practices

Building on this scholarship, in Spring of 2024, ADVANCE BU surveyed all Bradley chairs about how they evaluate teaching and their perceptions of the efficacy and equity of current evaluation methods. We had a response rate of 72%, with chairs from all five colleges and the library responding. As reflected in the figures below, the majority of chairs utilize qualitative and numerical responses on SETs, and classroom observations (at least for pre-tenure faculty) to evaluate teaching. The majority (57%) of chairs also found the Faculty Handbook “not at all helpful” or only “slightly helpful” when evaluating teaching. Only 39% of chairs considered their methods of evaluation “very effective” and only 17% considered current methods of evaluation “very equitable.” Because superior teaching is the top priority in faculty evaluation, and because Bradley’s primary mission is providing a quality education, it is crucial to ensure that evaluations of teaching are both effective and equitable.









Due to the prominence of SETs in Bradley's current evaluation of teaching and the apparent need to improve on both the effectiveness and fairness of evaluations, an ADVANCE BU team of faculty from all five colleges and staff from CTEL and Learning Technology has completed an analysis of all questions used on SETs at Bradley. A total of 395 questions are currently in use. But because many units utilize the same questions, there are only 44 unique questions. Informed by the literature on SETs, the team sought to identify those questions that are most likely to yield reliable and equitable data, specifically: (1) questions that students have the necessary expertise to answer; (2) questions that focus primarily on observable instructor behaviors or student learning; (3) questions that provide additional context (for instance about student attendance and effort) that can aid with the interpretation of results; and (4) questions that minimize opportunities for abusive or discriminatory comments based on the instructor's personal characteristics unrelated to course delivery. After a robust review process, a set of proposed SET items was produced and is included at the end of this document. This set includes the most effective and equitable questions currently in use at Bradley (some slightly reworded in alignment with the scholarship on SETs) in addition to several questions recommended in the literature. The set of questions includes some that may be used as a component of summative evaluation, and some that should only be used for the purpose of formative evaluation (ongoing improvement), as noted below.

ADVANCE BU Recommendations

1. That our standardized course survey instruments be renamed Student Experience Questionnaires (hereafter SEQs) or similar terminology to better reflect the nature of the data.

2. That each unit formally articulate what it means by "effective teaching" and use this as a basis for annual evaluations. We cannot effectively measure what we cannot define. A number of units on campus have already integrated detailed articulations of teaching effectiveness into their T&P guidelines, and their documents can serve as models for others.

3. That our standardized survey instruments provide students with guidance on offering constructive feedback. For instance, "Your feedback will be used to improve this course. When providing written comments, please be specific (providing examples whenever possible); focus on observed course practices (rather than general characteristics of the instructor or the course, eg. "too strict" or "too hard."); and be respectful (abusive or derogatory comments based on race, gender, age, etc. are not appropriate or constructive)."

4. That the University as a whole adopt a list of common questions to be included on every evaluation to allow for more valid comparisons across the institution. While draft questions are included below, we recommend that the final list of questions be determined through campus discussion.

5. That each unit choose two or more “Context” questions, two or more “Course” questions, and two or more “Instructor” questions from the final menu of questions (from a drop-down menu in Watermark). Only “Course” and “Instructor” questions from this menu will be used to generate numerical averages to ensure greater comparability of numerical averages across the University.

6. That units, if desired, include customizable, discipline-specific questions for the purposes of formative evaluation (not to be factored into numerical averages).

7. That units adopt uniform, discipline-appropriate methods for administering the surveys to maximize response rates, with surveys normally being administered only to courses with 5 or more students. (The scholarship on SETs suggests a minimum class size of 10, both to safeguard student confidentiality and for greater statistical validity. While Bradley’s smaller class size may require a lower minimum, evaluators should be mindful of the limited statistical validity of small sample sizes.)

8. That the University consider adopting a review protocol to allow for the removal of abusive or discriminatory evaluations from the instructor’s record.

9. That units use student survey data as only one of several sources of evidence for the overall annual summative “score” for teaching, and adopt additional practices from the list of Recommended Complementary Evaluation Practices below.

Recommended Complementary Evaluation Practices

Scholarship on effective and equitable teaching evaluation recommends relying on SEQs for no more than 30-50% of a summative “score.” To complement results from SEQs, we recommend that units utilize one or more of the methods below to provide triangulating forms of evidence from multiple perspectives—students, chair and/or peers, and the instructor (Berk 2018). The ultimate goal is to have a detailed, multifaceted record of each faculty member’s teaching activities, and of their demonstrated efforts to continually improve and remain current with evolving pedagogical developments.

Annual teaching portfolio: One way to compile all forms of evidence in one place is an annual teaching portfolio, including all course syllabi; a selection of course assignments and other instructional materials; CRCRS/curriculum proposals; copies of teaching-related (Scholarship of Teaching and Learning, SoTL) publications and conference abstracts; and evidence of any research, service, or professional development that is related to teaching.

The Watermark platform for Faculty Activity Reports already provides fields for such information, including, among others:

- Describe any instruction innovations that you introduced into this course during the current year (e.g., international issues, computer applications, ethical analysis, new classroom techniques, etc.).
- Describe any new teaching material (e.g., cases, videotapes, audiotapes, course modules, instructor manuals, test banks, or simulations) that you developed and/or implemented.
- Describe any activities in your course that enhanced student learning and/or student contact with the business community (e.g., guest speaker, SBDC, SBI, or outside projects, field trips, field projects, etc.).
- Describe and track activities related to course coordination, development, or preparation activities.
- If applicable, please provide a brief description of any interdisciplinary activity.
- If applicable, please briefly describe any positive societal impact of the activity.
- Enter any experiential learning activities you led that aren’t already included in the Schedule of Classes.
- Use this section to track professional development, such as conference and workshop attendance, fellowships or internships, or self-study course.

Such prompts can not only help generate ideas for instructional enhancement, but when faculty are encouraged to document their activities for each course they teach each semester, this provides documentation of instructional efforts over time. This data can then be compared to data such as student course evaluations or averages of final course grades to demonstrate instructional improvement over time when the instructor goes up for tenure and promotion.

Even if units choose not to formalize annual teaching portfolios or more detailed Watermark documentation, a number of other practices can contribute to more meaningful and more equitable evaluations.

Classroom observations by chair or peers: The majority (71%) of Bradley chairs report that they already utilize classroom observations (at least of pre-tenure faculty), a practice recommended in the literature. There is scholarly consensus, however, that to improve teaching effectiveness, classroom observations should be formative rather than summative (Arend 2018). For this reason, it is advisable to have a peer who is not involved in summative evaluation provide observation feedback to the instructor directly, preferably using a standardized observation form/rubric. Ideally, a team of trained faculty Teaching Fellows supported by CTEL (or at least a faculty member who does not vote on the tenure or promotion of the person they are observing) would provide this formative feedback. If a unit wishes to incorporate classroom observations into summative evaluations (annual "scores") as well, this should be done separately by another departmental peer or the chair, using a standardized observation form/rubric. Nancy Van Note Chism's (2007) *Peer Review of Teaching: A Sourcebook* provides a variety of forms for this purpose.

Review of syllabus and course materials by chair or peers: This might include a review of selected textbooks, course policies, the weekly course schedule, assignment instructions, tests or other assessments, point distributions, and samples of student work. The purpose of this review would be for peers to offer concrete suggestions to improve instructional effectiveness. Such a review could inform summative evaluation scores.

Measurements of student learning: This might include, among other possibilities, scores on a standardized exam that all students must take. It might include course-embedded assessments tied to key course objectives. It might also include a review of student work ("artifacts"). Such a review could inform summative evaluation scores.

Other professional development activities: A majority (79%) of Bradley chairs report factoring "professional development activities related to teaching" into their summative evaluations. Many of these activities (such as participation in teaching workshops, conferences, and courses) can already be documented through Watermark. Additionally, we recommend that CTEL establish a system to incentivize and document peer observation of exemplary teaching. Regardless of subject area or academic discipline, good teaching is good teaching, and by observing others, faculty gain ideas for instructional strategies such as beginning and ending class, introducing/transitioning to a new topic, pacing, questioning, wait time, movement around the classroom, student engagement/responses, explicit connections between instructional objectives and class activities/course content, use of instructional technology, and instructor movement around the classroom. Instructors have many different styles, and what works for one instructor might not work for another. Therefore, observing different instructors from various disciplines or subject areas is recommended. Participation in these documented activities could inform summative evaluation scores.

Proposed Student Experience Questionnaire Items

Context Questions

1. On average, how many hours per week did you spend outside of the class doing readings, reviewing notes, and any other related work for this course? [Not included in instructor's numerical average]
   a. 0-2
   b. 3-4
   c. 5-6

   d. 7-8
   e. 9-10
   f. 11+

2. How many absences have you had in this course? [Not included in instructor's numerical average]
   a. 0
   b. 1
   c. 2
   d. 3
   e. 4+

3. What, if anything, might you have done differently to be more successful in this class? [Open response. Not included in instructor's numerical average.]

Course Questions

4. The course syllabus or Canvas site provided clear and detailed information about course objectives, schedules, assignments, and policies (about grading, attendance, class participation, etc.).
   a. Strongly agree
   b. Somewhat agree
   c. Neither agree nor disagree
   d. Somewhat disagree
   e. Strongly disagree

5. The course enabled me to acquire new knowledge or skills and/or to reconsider my understanding of the subject.
   a. Strongly agree
   b. Somewhat agree
   c. Neither agree nor disagree
   d. Somewhat disagree
   e. Strongly disagree

6. The course structure, content, and presentations were clear and well organized.
   a. Strongly agree
   b. Somewhat agree
   c. Neither agree nor disagree
   d. Somewhat disagree
   e. Strongly disagree

7. Course activities, assignments, and assessed work corresponded closely to course materials and objectives.
   a. Strongly agree
   b. Somewhat agree
   c. Neither agree nor disagree
   d. Somewhat disagree
   e. Strongly disagree

8. What were the three (or more) most valuable concepts or skills that you gained from the course? [Open response, for formative evaluation.]

9. What aspects of this course would you suggest changing in the future to improve student learning? Check all that apply. [For formative evaluation. Not included in instructor’s numerical average.]
    a. The syllabus or Canvas site
    b. Course materials (textbook, readings, manuals, PowerPoints, etc.)
    c. In-class activities (lectures, discussions, group work, etc.)
    d. Tests and examinations
    e. Assignments
    f. Grading
    g. Instructor's preparation for each class period
    h. Instructor's knowledge of the subject
    i. Instructor's responsiveness to students
    j. Other, please specify [Open response]

<u>Instructor Questions</u>

10. Class sessions were engaging and contributed significantly to my learning.
    a. Strongly agree
    b. Somewhat agree
    c. Neither agree nor disagree
    d. Somewhat disagree
    e. Strongly disagree

11. The instructor treated students with respect and fostered an environment where I felt comfortable sharing my ideas.
    a. Strongly agree
    b. Somewhat agree
    c. Neither agree nor disagree
    d. Somewhat disagree
    e. Strongly disagree

12. The instructor provided timely and constructive feedback of my work.
    a. Strongly agree
    b. Somewhat agree
    c. Neither agree nor disagree
    d. Somewhat disagree
    e. Strongly disagree

13. Did you ever reach out to the instructor outside of class with a question or concern? [Exclude this question from numerical averages]
    a. No [Use skip logic to skip next question]
    b. Yes [Use skip logic to take “Yes” answers to next question]

14. The instructor was available to meet with students or respond to student questions or concerns outside of class.
    a. Strongly agree
    b. Somewhat agree
    c. Neither agree nor disagree
    d. Somewhat disagree
    e. Strongly disagree

15. I would recommend this instructor to other students.
    a. Strongly agree
    b. Somewhat agree
    c. Neither agree nor disagree
    d. Somewhat disagree
    e. Strongly disagree

16. Which of the following contributed positively to your learning in the class? Check all that apply. [For formative evaluation. Not included in numerical averages]
    a. Lectures and presentations
    b. Discussions, group work, and other classroom activities
    c. Quality of texts and other instructional materials (worksheets, manuals, PowerPoints, videos)
    d. Assignments
    e. Instructor's enthusiasm
    f. Instructor’s responsiveness to student questions, concerns, and needs
    g. Other (Please specify) [Open response]

17. What suggestions do you have for changes that you think would improve student learning in the class? [Open-response. For formative evaluation.]

References Cited

Arend, B. (2018) Towards a Comprehensive Teaching Evaluation Framework, University of Denver. Retrieved from https://otl.du.edu/wp-content/uploads/2021/10/Towards-a-Comprehensive-Teaching-Evaluation-Framework.pdf.

Arreola, R. A. (2000). *Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system.* Bolton, MA: Anker Publishing Co.

Benton, S. L., & Ryalls, K. R. (2016). *Challenging Misconceptions About Student Ratings of Instruction. IDEA Paper #58.* Manhattan, KS: The IDEA Center. Retrieved from https://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA_58.pdf.

Berk, R. A. (2018). Start spreading the news: Use multiple sources of evidence to evaluate teaching. *Journal of Faculty Development, 32*(1), 73-81.

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.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics*, 145, 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006

Boring, A., Ottoboni, K., and Stark, P. B. (2016, February 4). Student evaluations of teaching are not only unreliable, they are significantly biased against female instructors [Online resource]. Impact of Social Sciences Blog; The London School of Economics and Political Science. http://blogs.lse.ac.uk/impactofsocialsciences/

Braga, M., Paccagnella, M., and Pellizzari, M. (2014). Evaluating students’ evaluations of professors. *Economics of Education Review*, *41*(C), 71–88.

Chávez, K., and Mitchell, K. M. W. (2020). Exploring bias in student evaluations: Gender, race, and ethnicity. *PS: Political Science & Politics*, 53(2), 270–274. https://doi.org/10.1017/S104909651 9001744

Esarey, J., & Valdes, N. (2020). Unbiased, reliable and valid evaluations can still be unfair. *Assessment and Evaluation in Higher Education.* https://doi.org/10.1080/02602938.2020.1724875

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, *4*(1), 1304016. https://doi.org/10.1080/2331186X .2017.1304016

Hoyt, D. P., & Pallett, W. H. (1999). *Appraising Teaching Effectiveness: Beyond Student Ratings. IDEA Paper #36.* Manhattan, KS: The IDEA Center. Retrieved from http://www.theideacenter.org/sites/default/files/Idea_Paper_36.pdf.

Langbein, L. (2008). Management by results: Student evaluation of faculty teaching and the mis-measurement of performance. *Economics of Education Review*, *27*(4), 417–428.

Stark P. B. and Freishtat R. (2014). An evaluation of course evaluations. *ScienceOpen Research. 0*(0):1-7. DOI: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Van Note Chism, N. (2007). *Peer Review of Teaching: A Sourcebook, 2nd Edition*. Bolton, MA: Anker Publishing Co.

Weaver, Gabriela C. Ann E. Austin, Andrea Follmer Greenhoot & Noah D. Finkelstein (2020) Establishing a Better Approach for Evaluating Teaching: The TEval Project. *Change: The Magazine of Higher Learning 52*(3), 25-31, DOI: 10.1080/00091383.2020.1745575

Wines, W. A., and Lau, T. J. (2006). Observations on the folly of using student evaluations of college teaching for faculty evaluation, pay, and retention decision and its implications for academic freedom. *William & Mary Journal of Women and Law, 13*(1), 167-202