Interpreting Student Evaluation of Teaching (SET) Results: Guidelines for Deans, Department Heads, and Faculty

Prepared by Faculty Standards Committee (FSC), April 4, 2022

Approved by the University Senate May 2, 2022

In March 2010, the University Senate passed a motion endorsing the use of student evaluations of teaching (SETs), recognizing that they provide information on the student perception of their learning experience that can be useful for improving teaching. However, the Senate also urged caution in interpreting numerical values from SETs as an indicator of teaching competence. This caution is based on three premises:

1. As explicitly recognized by the Senate, no set of numerical values suffices as the sole indicator of teaching effectiveness. The collective bargaining agreement between the University of Connecticut Board of Trustees and the AAUP explicitly prohibits reliance on SETs as the only evidence of teaching effectiveness.
2. Although an overall score on an individual teaching evaluation can be an indicator of teaching performance, research shows that SET results are only moderately correlated with teaching effectiveness and can be influenced by factors that are not under the control of the instructor and are unrelated to teaching performance, such as student bias.
3. SETs are student ratings intended to represent the collective views of a group of students who have experienced the learning environment created by a faculty member. Student ratings are not a measure of student learning.

When used as one element in performance evaluations, SETs can have significant consequences for the careers of both full-time and part-time instructors. Thus, it is imperative that they be interpreted carefully. Given changes in student attitudes and expectations over time, as well as pedagogical methods used by faculty, the Faculty Standards Committee recommends regularly (every 5 years) revisiting the SET survey and guidelines to ensure they are up-to-date and reflective of current thinking and best practices. This process should include a review of the SET survey item, response scales, survey format, and mode of survey recruitment and administration. The following is guidance on the interpretation of SETs as of AY 2021/2022. A university-wide task force to operationalize "evidence of teaching excellence beyond SET" (formerly known as SET+ or SET plus) is being formed for AY 2022/2023.

Overall recommendation: In addition to considering the information provided by SETs, Deans, Department Heads, and PTR committee faculty are contractually obligated to use additional methods of evaluating instructors. All methods, to the extent they are contributing to the evaluation process, should be documented and agreed upon by the faculty in the department, transparent to those being evaluated, and collected from different independent lines of evidence (information sources).

Factors other than teaching competence that can influence SET results: The literature on SETs is both extensive and complex (see the appendix for a partial list of references). Although it is difficult to isolate individual factors, research suggests that SET responses can be influenced by multiple and often intersecting biases, including the following (note – this is not an exhaustive list):

- *Student year*: First-year students tend to give the lowest ratings, graduate students the highest.
- *Course level*: Students tend to give lower ratings in required courses than in electives.
- *Course topic*: Students may rate instructors lower when the instructor's perceived view on controversial or uncomfortable topics are contrary to their own.
- *Instructor race or ethnicity*: Students sometimes give faculty identifying as BIPOC (Black, Indigenous and People of Color) or identifying as a cultural minority lower ratings.
- *Instructor's primary language*: Students sometimes give lower ratings to instructors who are non-native English speakers, speak with an accent, or don't use what is currently referred to as Standard American English.
- *Disciplinary culture*: Students sometimes give lower ratings to women in male-dominated disciplines such as science, mathematics, economics, engineering, and philosophy, or to men in female-dominated disciplines such as nursing independent of their competency.
- *Gender/sex*: Students can rate faculty lower who do not conform to heterosexual, gender binary and cisgender norms; students can also give lower ratings to faculty based on perceived gender or sex, regardless of actual competence (ie, male students rating female instructors lower).
- *Field of study/discipline:* Classes in sciences and engineering tend to receive lower ratings than those in the humanities.
- *Age:* Students may rate younger instructors lower than older instructors.

Note: Although some anecdotal evidence and popular belief suggest that SET results are correlated with expected grades (with easy graders receiving higher scores), this claim is not supported by systematic research. Rather, evidence shows that there is a strong correlation between instructor ratings and students' perception of learning outcomes.

Guidelines for Interpreting SET results:

Based on research related to SETs, the Senate recommends the following guidelines be used in interpreting SET results.

Individuals or committees entrusted with reviewing files for tenure, promotion, hiring, contract renewals, teaching awards, or other university purposes making use of the SETs should be familiar with SET interpretation guidance. For personnel or promotion decisions in particular, efforts should be made to assess whether numerical scores correspond with other sources of information for teaching evaluation (eg, peer reviews, substantive qualitative comments from students, instructor self-reflections and teaching statements, and other relevant information).

1. Examine the patterns of instructor ratings across time. Compare multiple and similar courses across multiple semesters to form generalizations about student perceptions of teaching effectiveness. Don't focus on outliers.
2. Avoid comparing the raw SET scores between instructors without any context.
3. Remember that the sample is not random and therefore may not be representative of the entire class.
4. Do not over-interpret small differences in median ratings. Variance is normal.
5. Do not use university-or departmental averages (means or medians) as a line separating "failing" and "passing" teaching performance; as noted above, SETs can vary significantly across disciplines, so comparing to university-wide averages may not be appropriate or informative.
6. Do not average multiple, inherently-different SET items into a single value. Composite scores can misrepresent data.
7. Ask: Are one or two low student ratings affecting the results in a small class?
8. Ask: Does this instructor receive consistently better ratings for some skills than others (preparation, clear assignments, receptivity to students)?
9. Ask: Are SET ratings influenced by large class size or courses outside of a student's major?
10. Ask: Is the distribution of SET ratings in particular classes bi-modal, as sometimes occurs in classes that include controversial or politically-charged topics?
11. Do not solely focus on the two questions related to overall ratings of the instructor's teaching and the course – examine the scores holistically.
12. Recognize that when there are responses from small numbers of students, percentages or average ratings may not be meaningful or representative.
13. Ask: Are student ratings consistent with other sources of evidence?

Appendix

Selected Relevant Publications:

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

Arreola, R. A. (2007). *Developing a Comprehensive Faculty Evaluation System.* San Francisco: Jossey-Bass.

Cashin, W. (1999). Student Rating of Teaching: Uses and Misuses. In P. Seldin (Ed.), *Changing Practices in Evaluating Teaching: A Practical Guide to Faculty Performance and Promotion/Tenure Decisions* (pp. 25-44). Boston: Anker Publishing Company, Inc.

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

Cohen, P.A. (1990). Bring research into practice. In M. Theall, & J. Franklin (Eds.), *Student Ratings of instruction: Issues for Improving Practice: New Directions for Teaching and Learning,* No. 43 (pp. 123-132). San Francisco: Jossey-Bass.

Esarey, J., & Valdes, N. (2020) Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education,* 45, p. 1106-1120.

Feldman, K. A. (1993). College students' views of male and female faculty college teachers: Part II – Evidence from students' Evaluations of their classroom teachers. *Resch in Higher Ed*, *34*, 151-211.

Hammermesh, D., Parker, A. (2005) Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review*, 24, p. 369-376.

Hendrix, K. G. (1998). Student perceptions of the influence of race on professor credibility. *Journal of Black Studies*, *28*, 738-764.

Hollman, M., Key, E., and Kreitzer, R. (2019) "Evidence of Bias in Standard Evaluations of Teaching"; a maintained bibliography of relevant studies last accessed 3/25/2022 https://docs.google.com/document/d/14JiF-fT--F3Qaefjv2jMRFRWUS8TaaT9JjbYke1fgxE/edit

Houston, T. Empirical Research on the Impact of Race & Gender in the Evaluation of Teaching. Report, Center for Excellence in Teaching & Learning. Seattle University. October 5, 2005

Houston, T. Research Report: Race and Gender Bias in Student Evaluations of Teaching. Center for Excellence in Teaching & Learning. Seattle University. October 31, 2005

Ory, J.C. & Ryan, K. (2001). How do student ratings measure up to a new validity framework? In M. Theall, P.C. Abrami, & L.A. Mets (Eds.), *New Directions for institutional research: No. 109. The student ratings debate: Are they valid?* San Francisco: Jossey-Bass.

Rubin, D. L. (1998). Help! My professor (or doctor or boss) doesn't talk English. In J. N. Martin, T. K. Nakayama, L. A. Flores (Eds.), *Readings in Cultural Contexts* (pp. 149 – 160). Mountain View, CA: Mayfield Publishing Company.

Seldin, P. (1999) Changing practices in Evaluating Teaching. Bolton, MA. Anker.

Uttl, B., White, C.A., and Gonzalez D.W. (2017) Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation,* 54, p.22-42