Report to the Faculty Senate

From: Ad Hoc Committee on Student Evaluations of Faculty Teaching

Date: April 10, 2024

The committee has reviewed relevant literature, SU's current practices, and considered best practices with respect to the role of Student Evaluations of Teaching (hereafter SET) for evaluating faculty for Tenure and Promotion. The evidence is that SET provide little, if any, valuable information about faculty teaching and student learning because SET register a wide range of student biases regarding discipline, race, gender, physical attractiveness, likeability, and ease of grading. (See appendices for details and an extensive review of relevant literature).

Below are our recommendations.

1. Faculty are expected to solicit student feedback on classroom experiences, particularly in the form of mid-semester formative evaluations, and use that information to guide adjustments in their courses. (See Appendix B).
2. Faculty are not expected/required to share the raw data of SET with supervisors or as part of Tenure or Promotion packets.
    a. Faculty must continue to report on their responses to SET in their annual self-evaluation; inclusion of raw data in those reports is at the discretion of the individual faculty member.
    b. Departments and schools are encouraged to use other practices for evaluating teaching that serve their programs, for example peer evaluations. GULLWEEK, as well as our courses, allow us to assess student learning.
3. Faculty will develop and maintain a teaching portfolio for the purposes of evaluation of their teaching for Tenure and Promotion.
    a. Teaching Portfolios will include relevant teaching materials, such as syllabi, assignments, sample student work, and other, as well as self-evaluation materials.
    b. Departments/schools will develop guidelines for teaching portfolio contents and structure (See Appendix C)
4. Full-time, tenure-track, faculty will participate in ongoing, small, interdisciplinary support groups for faculty development in teaching excellence as part of their progress toward Tenure. Schools, departments, and participating faculty will have discretion regarding meeting times and regularity to be NOT LESS THAN once per semester.
    a. Effective faculty development support groups for teaching excellence will be facilitated by experienced faculty with good teaching records, deep familiarity

with the University's mission and General Education Program, and desire to support junior faculty.

b. Optimal size for groups is no larger than 6 (7 with facilitator).

5. As part of the annual review process, supervisors are expected to engage in regular face-to-face meetings with junior faculty to discuss effective teaching strategies and practices.

Ad Hoc Committee on Student Evaluations of Faculty Teaching

Committee members:

- Heather Matthews;
- Michele Schlehofer;
- Alison Dewald;
- Anne Chillingworth-Shaffer;
- April Logan;
- Memo Diriker;
- Shan Lei;
- Shawn McEntee;
- Starlin Weaver

Senate Rep: Memo Diriker

Charge: The Ad Hoc Committee on Student Evaluations of Faculty Teaching shall consider the role of student evaluations in faculty evaluations and in promotion and tenure decisions at SU. The committee shall look at the literature on best practices for student evaluations of faculty teaching and on the suitability and equity of using student evaluations as an indicator of teaching quality. The committee shall also reach out to all SU faculty for feedback on the matter. The committee shall create a report that summarizes its findings and provides a concrete recommendation for how (or if) student evaluations should be used in faculty evaluations and in promotion and tenure decisions at SU; the report shall include a draft of any proposed changes to existing policies. The report shall be presented to the Faculty Senate no later than the first Faculty Senate meeting in Spring 2024 and submitted to the Faculty Senate President at least one week in advance of said meeting.

Appendix A: SET Literature Summary

For every complex problem there is an answer that is clear, simple, and wrong." H. L. Mencken (Quoted in Uttl 2017)

The use of student evaluations of teaching (SET) to measure teaching effectiveness is common, as SET are convenient and inexpensive to administer and allow students to have a voice in evaluating their faculty's teaching. Students are uniquely positioned to report on their class experiences and perceptions, and SET provide feedback as well as simple numerical comparisons for administrators to use for instructor assessment and evaluation (Stark & Freishtat, 2014; Uttl 2017). However, a large body of research demonstrates that SET do not measure teaching effectiveness nor student learning and are instead influenced by factors such as the instructor's discipline, likeability, ease of grading, gender, race, and physical attractiveness.

If SET were representative of effective teaching, then we would expect them to correlate positively with student achievement, which may be measured as scores on assessments across different sections or student success in subsequent related courses. Studies investigating this relationship have found either no or negative correlation (Johnson, 2003; for reviews, see Kornell & Hausman, 2016; Carpenter 2020). For example:

- Studies in which students were randomly assigned to various instructors in calculus, economics, management, or law courses found either no correlation or negative correlation between SET and scores on standardized tests of the material (Braga et al., 2014; Carrell & West, 2010).

- When controlling for GPA and ACT scores, Yunker & Yunker found a negative relationship between students' course evaluations in Introductory Accounting and students' later performance in Intermediate Accounting and Weinberg, Hashimoto, & Fleisher showed in large sample studies that students taught by highly rated professors in prerequisites performed worse in follow-up courses (Weinberg et al., 2007; Yunker & Yunker, 2003).

The best designed studies of effective instruction follow multi-section courses for which: 1) there are many course sections with the same material taught, 2) students are randomly assigned to different sections and/or the study controls for prior student learning/ability, 3) all sections are assessed with the same centrally administered exam. In 1981 Cohen et al. published a highly cited meta-analysis of multi-section studies and reported a small to moderate ($r = 0.43$) correlation between SET scores and student achievement (Cohen, 1981). However, subsequent analyses of this data and subsequent data suggest that the correlation was an artifact of over-valuing small sample-size studies and publication bias, wherein studies reporting statistically significant correlations are more likely to be published (Clayson, 2009; Stroebe, 2020; Uttyl, 2017). Since then,

- A more recent meta-analysis of nearly 100 multi-section studies indicates that SET /learning correlation is small (r = 0.12). When prior student ability is considered, the correlation is zero (r = -0.06). (Uttle, 2017)
- A 2016 study of 23,000 SET scores from 4,423 first year students in 1,177 sections in France found the correlation between SET and final exam scores to be r= 0.04. Of note, SET were compulsory, so the student response rate was nearly 100%, and the students had been unable to self-select into different sections. (Boring, 2016).

SET are consistent, in that evaluations for a given instructor positively correlate within the same course and over time (Carpenter, 2020). This suggests that they reflect something stable about the instructor. Research points to a number of factors, including:

Course Discipline – Multiple studies indicate that SET are highest for courses in humanities and language and lowest in math, engineering, and science. A study of 238,471 classes found SET for courses in natural sciences to be 0.30 standard deviations lower than for those in humanities. Additional studies have identified engineering, computer science, and chemistry as the departments with lowest ranking SET. An analysis of SET at NYU found the average score in English to be 4.29; in math only 3.68 (Uttyl & Smibert, 2017).

Instructor Likeability – Carpenter et al (2016, 2013, 2018; reviewed in 2020) did a series of experiments for which students watched a video by an instructor using a fluent or disfluent style, then were asked to estimate how much they learned and to complete a test on the lecture content. The fluent instructor was rated as more organized and students judged their learning to be higher, though test scores between groups were not significantly different. In studies where the instructor was asked to teach a course with 'enthusiasm', students rated the instructor as more effective, better organized, having a higher level of knowledge, and using a higher quality textbook; they estimated having learned more, despite earning nearly identical grades.

This mirrors the "Dr. Fox Effect" experiment in which an actor was introduced as an expert on mathematics applied to human behavior. The actor was instructed to give a lecture intentionally meaningless, vague, and contradictory, but to do so with enthusiasm and passion. Audience feedback was overwhelmingly positive, with 90% reporting that the lecture was well-organized, interesting, and contained clear examples (Naftulin et al., 1973).

A study of 1486 students at the UC Berkley Department of Statistics found the correlation between instructor effectiveness and course enjoyment to be 0.75. (Stark, 2014)

Evaluations collected from students after 5 minutes of exposure to a professor accurately predicted the SET scores at the semester's end, underscoring the superficiality of the assessment (Clayson & Sheffet, 2006; similar result in Merritt, 2007).

Course Ease - Multiple studies indicate that instructors who grade leniently receive higher average ratings than those who do not (DuCette & Kenney, 1982; Eiszler, 2002; Ewing, 2012; Greenwald & Gillmore, 1997a, 1997b; Holmes, 1972; Isely &Singh, 2005; McPherson, 2006; Olivares, 2001; Stroebe, 2016; but see Heckert, Latier, Ringwald-Burton, & Drazen, 2006; Marsh & Roche, 2000; Palmer et al., 1978; all from Wolfgang 2020; Berezvai 2021). A 2014 study found that instructors' mean SET scores decreased after implementation of an anti-grade inflation policy (Butcher, McEwan, and Weerapana) and Rosen et al found correlations of 0.62-0.66 between overall course quality and 'easiness', though the correlation was weaker at highly competitive universities (Chiu et al 2019, Radchenko 2020).

Gender – Abundant evidence suggests that women receive systematically lower SET scores than men, though some studies did not find bias and others showed women receiving higher evaluations (Adams, 2022, Hoorens 2021, Joye 2015, Velencia 2022). Generally, women are penalized for not behaving in gender stereotypical ways while men are not. For reviews, see Andersen & Miller, 1997; Basoe & Martin, 2012, Carpenter 2020.

A large, multi-section study by Hamermesh and Parker (2005) found that female instructors received ratings averaging nearly half a standard deviation below those of their male counterparts.

In another large, multi-section study students were randomly assigned to either male or female professors. There was no correlation between instructor gender and student course grades, study time, or grade in the next course in sequence. However, female instructors received rankings 37% lower than those of the males, and the effect was strongest in courses with mathematical content and for junior instructors (Mengel 2019).

Similarly, in the study of 23,000 randomly assigned French students cited above, there was a statistically significant correlation between instructor gender and SET (with male instructors having higher SET, and greatest difference coming from male students) despite the students of male instructors scoring slightly lower on final exams (Boring Ottoboni Stark 2016),

In 2015 MacNell, Driscool, and Hunt had a male and female instructor each teach an online course with identical content. The instructors only communicated via email and message boards. Half of the male instructor's students thought the instructor was female and vice versa. Overall, there were no differences in student evaluation of teaching effectiveness according to the actual gender of their instructor. However, students gave more positive evaluations to the instructors they believe to be male, regardless of the instructor's actual gender. The male identity received significantly higher scores on professionalism, promptness, fairness, respectfulness, enthusiasm, and giving praise. (MacNell, et al 2015)

In 2003, Arbuckle and Williams had students view a 30-minute lecture with slides narrated by an age and gender-neutral voice. Afterwards, students were told that the reader was young and female, young and male, older and female, or older and male. They took a test on the lecture content and evaluated the presentation. Test scores did not differ between groups, however, students who believed the instructor was male gave significantly higher ratings of the reader's

enthusiasm and meaningful tone than those who were told the instructor was female. In addition, students told the instructor was young rated the narration higher than those told the instructor was old. Additional negative age correlation with SET are reported by Sprinkle 2008; Goebel & Cashen, 1979).

Ethnicity and Race - Evaluations of SET scores as a function of race are often confounded by small group size, and this area of research remains understudied. However, evidence shows that racial minority, nonwhite, and non-native-English-speaking instructors receive systematically lower SET ratings than their white/non-minority counterparts, even when factors such as course type or student average score are accounted for (Chávez 2020, Chisadza 2019). A study of 5630 faculty at liberal arts colleges found that racial minority faculty members were rated less favorably on quality, helpfulness, and clarity and more positively on easiness (Reid 2010). The impact is largest on black lecturers, with other minority faculty falling between their black and white colleagues (reviewed in Wolfgang, 2020 and Heffernan, 2022). In 2016, Sanchez & Khan demonstrated that a male instructor speaking with a non-English accent received lower ratings of instructional quality than a male instructor presenting identical content with no accent, although students' learning of the content was not affected (Sanchez & Khan, 2016).

Candy Distribution - Youmans & Jee had experimenters give chocolate candy to three of six discussion sections in a large course before students completed online course evaluations. The sections offered chocolate gave the course more positive evaluations than those who were not. (Youmans & Jee, 2007).

References

Adams, S., Bekker, S., Fan, Y., Gordon, T., Shepherd, L. J., Slavich, E., & Waters, D. (2022). Gender bias in student evaluations of teaching: 'Punish[ing] those who fail to do their render Right'. *Higher Education, 83,* 787–807. https://doi.org/10.1007/s10734-021-00704-9

Andersen, Kristi, and Elizabeth D. Miller. "Gender and student evaluations of teaching." *PS: Political Science & Politics*, vol. 30, no. 2, June 1997, pp. 216+. *Gale Academic OneFile*, link.gale.com/apps/doc/A287748409/AONE?u=umd_ub&sid=googleScholar&xid=96187ed3. Accessed 10 Apr. 2024.

Arbuckle, J., & Williams, B. D. (2003). Students' Perceptions of Expressiveness: Age and Gender Effects on Teacher Evaluations. *Sex Roles: A Journal of Research, 49*(9-10), 507–516. https://doi.org/10.1023/A:1025832707002

Berezvai, Z., Dániel Lukáts, G., & Molontay, R. (2021). Can professors buy better evaluation with lenient grading? The effect of grade inflation on student evaluation of teaching. *Assessment & Evaluation in Higher Education, 46*(5), 793-808. http://doi.org/10.1080/02602938.2020.1821866

Basow, S.A. and Martin, J.L. (2012) Bias in Student Evaluations. In: Kite, M.E., Ed., *Effective Evaluation of Teaching: A Guide for Faculty and Administrators*, Society for the Teaching of Psychology, Washington DC, 40-49.

Boring, A., Ottoboni, K., & Stark, P. S. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research,* 0(0):1-11. http://doi/org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Braga, M., Paccagnella, M., & Pellizzari, M. Evaluating students' evaluations of professors. *Economics of Education Review*, 41, 71-88. https://doi.org/10.1016/j.econedurev.2014.04.002

Butcher, K.F., McEwan, P.J., & Weerapana, A. (2014). THe effects of an anti-grade-inflation policy at Wellesley College. *Journal of Economic Perspectives*, 28(3), 189-204.

Carpenter, S. K., Witherby, A. E., & Tauber, S. K. (2020). On students' (mis)judgments of learning and teaching effectiveness. *Journal of Applied Research in Memory and Cognition, 9*(2), 137–151. https://doi.org/10.1016/j.jarmac.2019.12.009

Carrell, S.E. & West, J.E. Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3), 409-132. https://doi.org/10.1086/653808

Chávez, K., & Mitchell, K. M. W. (2020). Exploring bias in student evaluations: Gender, race, and ethnicity. *PS: Political Science & Politics*, *53*(2), 270–274. http://doi.org/10.1017/S1049096519001744

Chisadza, C., Nicholls, N., & Yitbarek, E. (2019). Race and gender biases in student evaluations of teachers. *Economics Letters, 179,* 66-71. https://doi.org/10.1016/j.econlet.2019.03.022

Chiu, Y. Chen, K., Hsu, Y., & Want, J. (2018). Understanding the perceived quality of professors' teaching effectiveness in various disciplines: the moderating effects of teaching at top colleges. Assessment & Evaluation in Higher Education, 44(3), 449-462. DOI:10.1097/00001888-197307000-00003

Clayson, D. E., & Sheffet, M. J. (2006). Personality and the Student Evaluation of Teaching. *Journal of Marketing Education, 28*(2), 149–160. https://doi.org/10.1177/0273475306288402

Clayson, Dennis. (2009). Student Evaluations of Teaching: Are They Related to What Students Learn?. *Journal of Marketing Education.* 31. 16-30. 10.1177/0273475308324086.

Clayson, D. (2022). The student evaluation of teaching and likability: what the evaluations actually measure. *Assessment & Evaluation in Higher Education, 47*(2), 313-326. http://doi.org/10.1080/02602938.2021.1909702.

Cohen, P. A. (1981). Student Ratings of Instruction and Student Achievement: A Meta-analysis of Multisection Validity Studies. *Review of Educational Research*, 51(3), 281-309. https://doi.org/10.3102/00346543051003281

Esarey, J., & Valdes, N. (2020) Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment & Evaluation in Higher Education, 45*(8), 1106-1120. http://doi.org/10.1080/02602938.2020.1724875

Goebel, B. L., & Cashen, V. M. (1979). Age, sex, and attractiveness as factors in student ratings of teachers: A developmental study. Journal of Educational Psychology, 71(5), 646-653. http://dx.doi.org/10.1037/0022-0663.71.5.646

Hamermesh, D.S., & Parker, A. (2005). Beauty in the classroom: instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review*, 24(4), 169-376. https://doi.org/10.1016/j.econedurev.2004.07.013

Heffernan, T. (2022). Sexism, racism, prejudice, and bias: a literature review and synthesis of research surrounding student evaluations of courses and teaching, A*ssessment & Evaluation in Higher Education, 47*(1), 144-154. http://doi.org/10.1080/02602938.2021.1888075

Heffernan, T. (2023). Abusive comments in student evaluations of courses and teaching: the attacks women and marginalised academics endure. *Higher Education,* 85, 225–239. https://doi.org/10.1007/s10734-022-00831-x

Hoorens, V., Dekkers, G. & Deschrijver, E. (2021). Gender bias in student evaluations of teaching: Students' self-affirmation reduces the bias by lowering evaluations of male professors. *Sex Roles, 84*, 34–48. https://doi.org/10.1007/s11199-020-01148-8

Johnson, V. E. 2003. Grade Inflation: A Crisis in College Education. New York: Springer-Verlag

Joye, S., & Wilson, J. H. (2015). Professor age and gender affect student perceptions and grades. *Journal of the Scholarship of Teaching and Learning*, *15*(4), 126–138. https://doi.org/10.14434/josotl.v15i4.13466

Kornell, N. & Hausman, H. (2016). Do the best teachers get the best ratings? Front. Psychol., 7(570). https://doi.org/10.3389/fpsyg.2016.00570

MacNell, L., Driscoll, A. & Hunt, A.N. (2015). What's in a name: exposing gender bias in student ratings of teaching. *Innovations in Higher Education, 40,* 291–303. https://doi.org/10.1007/s10755-014-9313-4

Mengel, F., Sauermann,J., & Zölitz, U. (2019). Gender Bias in Teaching Evaluations, *Journal of the European Economic Association*, 17(2), 535–566. https://doi.org/10.1093/jeea/jvx057

Merritt, D. J. (2007). Bias, the brain, and student evaluations of teaching. O*hio State Public Law Working Paper No. 87.* https://ssrn.com/abstract=963196 or http://dx.doi.org/10.2139/ssrn.963196

Ratings of Teaching. *Innovations in Higher Education, 40,* 291-303. http://doi.org/10.1007/s10755-014-9313-4

Mengel, F., Sauermann, S., & Zölitz, U. (2019). Gender bias in teaching evaluations. *Journal of the European Economic Association*, *17*(2), 535–566, https://doi.org/10.1093/jeea/jvx057

Naftulin, D.H., Ware, J.E., & Donnelly, F.A. The Doctor Fox lecture: A paradigm of educational seduction. Journal of Medical Education 48(7): 630-5. DOI:10.1097/00001888-197307000-00003

Radchenko,N. (2020). Student evaluations of teaching: Unidimensionality, subjectivity, and biases. *Education Economics, 28*(6), 549-566. http://doi.org/10.1080/09645292.2020.1814997

Rosen, A. S. (2018). Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors.com data. *Assessment & Evaluation in Higher Education*, *43*(1), 31–44. https://doi.org/10.1080/02602938.2016.1276155

Sanchez, C. A., & Khan, S. (2016). Instructor accents in online education and their effect on learning and attitudes. *Journal of Computer Assisted Learning, 32*(5), 494–502. https://doi.org/10.1111/jcal.12149

Simonson, S. R., Earl, B., & Frary, M. (2022) Establishing a framework for assessing teaching effectiveness. C*ollege Teaching, 70*:2), 164-180. http://doi.org/10.1080/87567555.2021.1909528

Sprinkle, J.E. (2008). Student perceptions of effectivenss: an examination of the influence of student biases. *College Student Journal*, 42(2), 286-293.

Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. *Science Open Research, 0*(0),1-7. http://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Stroebe, W. (2020). Student evaluations of teaching encourages poor teaching and contributes to grade inflation: A theoretical and empirical analysis, *Basic and Applied Social Psychology, 42*(4), 276-294. http://doi.org/10.1080/01973533.2020.1756817

Uttl B, Smibert D. Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ.* 2017 May 9;5:e3299. doi: 10.7717/peerj.3299.

Uttl, B., White, C. A., & Wong Gonzalez, D. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54,* 22-42. https://doi.org/10.1016/j.stueduc.2016.08.007

Valencia, E. (2022). Gender-biased evaluation or actual differences? Fairness in the evaluation of faculty teaching. *Higher Education*,83, 1315–1333. https://doi.org/10.1007/s10734-021-00744-1

Weinberg, B.A., Fleisher, B.M., & Hashimoto, M. (2007). Evaluating methods for evaluating instruction: The case of higher education. *National Bureau of Economic Research Working Paper Series*, Working Paper 12844. **DOI** 10.3386/w12844

Youmans, R. J., & Jee, B. D. (2007). Fudging the numbers: Distributing chocolate influences student evaluations of an undergraduate course. *Teaching of Psychology, 34*(4), 245-247. https://doi.org/10.1080/00986280701700318

Yunker, P. J., & Yunker, J. A. (2003). Are Student Evaluations of Teaching Valid? Evidence From an Analytical Business Core Course. *Journal of Education for Business*, *78*(6), 313–317. https://doi.org/10.1080/08832320309598619

Appendix B: Alternatives to SET Literature Summary, SET Best Practices, and Recommendations

Alternatives to traditional student/course evaluations in evaluating faculty teaching

I. Mitigating Student Evaluation Influence

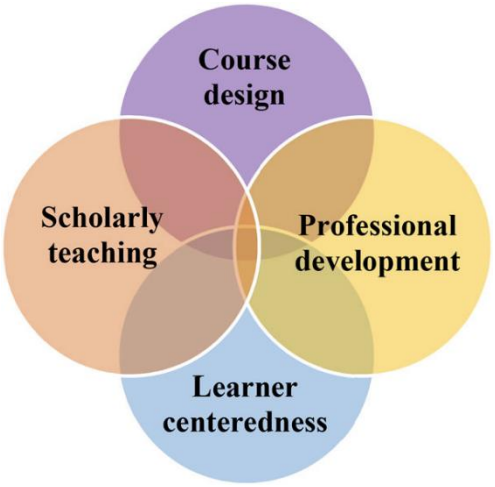
Studies concur on the importance of student evaluation. However, the evaluation of teaching is a complex process. Student evaluations may be used for (SE) "formative" decisions, "which use [SE] evidence to improve and shape the quality of our teaching, and/or "summative"

ones, “ which use [SE] evidence to ‘sum up’ our overall performance or status to decide about our annual merit pay, promotion, and tenure. However, SES is the single type evidence the SU faculty handbook requires faculty to include in a tenure and promotion dossier. In addition, many departments require faculty to include or reference SES as part of the annual review process. It is also problematic to privilege SES because of the potential biases that influence them; these are summarized by another subcommittee. Therefore, SU should not privilege SE as of the most important tools for evaluating teaching (Berk, 2005; Carpenter, Witherby, & Tauber, 2020; Simonson, Earl, & Frary, 2022; Shao, Anderson, & Newsome, 2007; Stark-Wroblewski, Ahlering, & Brill, 2007). Because it does not seem feasible for SU to abandon SES altogether, this subcommittee’s summary will focus on literature that offers strategies for 1) reducing the weight of SES or, at least, better balancing it with other possible forms of evidence for evaluating teaching (e.g., Hornstein, 2017) and 2) gathering student evaluation data that better reflects the diverse uses of SE. Studies recommend some common evaluation tools that might serve as alternatives or supplements to student evaluations: self-evaluation, administration evaluation, peer evaluation, teaching portfolio, student interviews, classroom visits, alumni ratings, employers' ratings (and job performance data), scholarship activities (both in teaching and in faculty's field of study), teaching awards; students learning outcome measures; course features (required or elective; grad/undergrad). In particular, Miller and Seldin (2014) conduct a study by collecting information through questionnaires to the academic deans of a random sample of accredited four-year liberal arts colleges. "Of 538 academic deans surveyed, 410 (76 percent) responded". Their survey describes types of evidence frequently used with student evaluation to create a picture of e faculty teaching performance: **chairs and deans’ evaluation, self-evaluation, classroom visits, faculty committee’s evaluation, research and publication (a perspective to show teaching competence), and an analysis of a professor’s course syllabi and examinations**. The complete results are shown in Table 2 of Miller and Seldin’s study. The SU Faculty handbook

The handbook states that student evaluations be part of evidence included in the tenure and promotion application. But it does not provide guidelines regarding how to interpret and weigh student evaluations. The explicit requirement of student evaluations implicitly privileges them over other types of evidence. It is recommended that inclusion of student evaluation be optional rather than required evidence in the tenure and promotion application.

II.Rethinking Faculty Evaluation Methods

Simonson, Earl, and Frary (2022) suggest that the evaluation of faculty teaching needs to involve methods that also make possible the assessment of student learning. They consider the complexity of **teaching and learning**, developed a four-element teaching effectiveness tool, shown in the following figure:



Source: Figure 2. Simonson, S. R., Earl, B., & Frary, M. (2022). Establishing a framework for assessing teaching effectiveness. *College Teaching*, *70*(2), 164-180.

They too develop a rubric and suggest a comprehensive approach to the types of evidence that can be used to demonstrate student learning as well as other elements of teaching effectiveness. They suggest using **syllabi, course assignments, student work samples, and course design table**s to assess course design. They suggest teachers implement evidence-based practices to demonstrate scholarly teaching, which can be assessed using, for example, **peer evaluation, and class observation**. Similarly, **syllabi, course assignments, and peer observation** can be used to assess if the teaching uses a learner-centered approach. They also point out to use of, for example, **mid-term survey**, **and reflection on course evaluation** as evidence for continuous teaching improvement. A complete rubric is shown in Figure 3 in this research. Perhaps faculty handbook and department T & P policies should provide a more detailed list of suggested items for demonstrating teaching effectiveness to ensure that faculty consider including materials that speak to course design and learning.

Like Simonson, Earl, and Frary, several other researchers asserted that the demonstration and discussion of learning outcomes might provide more accurate information regarding a faculty member's teaching effectiveness. To this end, Anders proposed using **focus groups and role-play** to solicit more candid and detailed reflections from students about their learning in a course. Borch, Sandvoll, and Risor advocated a similar type of tool by suggesting that faculty collaborate with students to create **"dialogue-based evaluation methods."** However, the Borch, Sandvoll, and Risor researched was conducted in Norway, and their proposed method raises important questions regarding the resources needed to execute it. Lastly, Stark-Wroblewski, Ahlering, and Brill, suggest that faculty conduct **pre- and post-assessments of students' knowledge of a course-related topic** to measure student learning.

III. Peer Observations Concerns

Although several of the aforementioned studies assert that peer statements/observations and class room visits can help mitigate the bias and other problems with student evaluations, Berk reveals that most faculty are resistant to them because their potential for bias, unfairness, and inaccuracy. Indeed, there is "consensus" in academia that "peer observation data should be used for formative [or developmental] rather than summative decisions." Yet, many departments at SU require faculty to include peer observations in their tenure and promotion application. Indeed, J.M. Golding and Philipp Kraemer question whether peer observations can infringe on academic freedom; therefore, our subcommittee might request that the Senate Academic Freedom and Tenure Committee explore this concern.

SET Design Best Practices

Given the many biases in student evaluation and often flawed interpretation of it, some studies provided suggestions for improving teaching evaluation design, such as 1) dropping the questions beyond students' capability; 2) drop the obscure questions, such as overall teaching effectiveness; 3) avoid comparing averages of teaching evaluation scores; 4) avoid comparing different courses at different course level and features (e.g., Hornstein, 2017). Specifically, Carpenter, Witherby, and Tauber (2020) propose to **develop a well-designed student evaluation** to mitigate the biases: 1) eliminating the evaluation questions beyond students' knowledge and capability (e.g., evaluate the professor's knowledge in the field of study); 2) relying more on students' qualitative comments but the information users need to get trained when interpreting

this information; 3) completing evaluations at multiple times throughout the semester to limit the negative effect of faulty memory be the end of semester. **They also admit that all these methods may not be able to solve the biases with student evaluations.** They suggest the following alternatives: peer evaluation/observation; student interview by administrators; teaching portfolio including one's teaching philosophy, syllabi, example lessons, assignments, and grading rubric; follow-up assessment about students' learning outcomes (e.g., performance in later courses)

Summary of Recommendations

- SU handbook should list a variety of evidence for faculty to choose from, rather than requiring specific types.
- Types of evidence should include but not be limited to:
    - self-evaluation,
    - administration evaluation,
    - faculty committee's evaluation
    - peer evaluation,
    - teaching portfolio,
    - student interviews,
    - classroom visits,
    - alumni ratings,
    - employers' ratings (and job performance data),
    - scholarship activities (both in teaching and in faculty's field of study),
    - publication (a perspective to show teaching competence),
    - teaching awards;
    - students learning outcome measures;
    - course features (required or elective; grad/undergrad),
    - and an analysis of a professor's course syllabi and examinations.
- Faculty, especially evaluators of student evaluations need to be trained about the interpretation of them
- Student evaluation should not be a required type of evidence
- Peer evaluation should not be a required type of evidence
- SU should continue exploring alternatives to using student evaluations to solicitate students feedback, such as focus group and role play, student interview, and pre-post assessment of student knowledge
- Develop a well-designed student evaluation to mitigate the biases: 1) eliminating the evaluation questions beyond students' knowledge and capability (e.g., evaluate the professor's knowledge in the field of study); 2) relying more on students' qualitative

comments but the information users need to get trained when interpreting this information; 3) completing evaluations at multiple times throughout the semester to limit the negative effect of faulty memory be the end of semester.

Bibliography

Anders, B. A. (2023, May). Using a Focus Group to Enhance Course Evaluation Inclusion and Feedback. In *Erasmus Scientific Days 2022 (ESD 2022)* (pp. 332-338). Atlantis Press.

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.

Bork, I., Sandvoll, R. & Risor T. (2020). Discrepancies in purposes of student course evaluations: what does it mean to be 'satisfied'? *Educational Assessment, Evaluation and Accountability*, 32, 83-102. Carpenter, S. K., Witherby, A. E., & Tauber, S. K. (2020). On students'(mis) judgments of learning and teaching effectiveness. *Journal of Applied research in memory and cognition*, *9*(2), 137-151.

Golding, J.M. & Kraemer, P.J. Observations of Professors: Tread Lightly. *Inside Higher Education*. July 23 2015. Website.

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, *4*(1), 1304016.

Miller, J. E., & Seldin, P. (2014). Changing practices in faculty evaluation. *Academe*, *100*(3), 35-38.

Simonson, S. R., Earl, B., & Frary, M. (2022). Establishing a framework for assessing teaching effectiveness. *College Teaching*, *70*(2), 164-180.

Shao, L. P., Anderson, L. P., & Newsome, M. (2007). Evaluating teaching effectiveness: Where we are and where we should be. *Assessment & Evaluation in Higher Education*, *32*(3), 355-371.

Stark-Wroblewski, K., Ahlering, R. F., & Brill, F. M. (2007). Toward a more comprehensive approach to evaluating teaching effectiveness: Supplementing student evaluations of

teaching with pre–post learning measures. *Assessment & Evaluation in Higher Education*, *32*(4), 403-415.

Appendix C: Best Practices – Teaching Portfolios

Washington State University:

Teaching Portfolio | Office of the Provost | Washington State University (wsu.edu)

Examples of Teaching Portfolios

Examples of Teaching Portfolios | Faculty | Washington State University (wsu.edu)

DePaul University – Teaching Commons

Teaching Portfolios | Reflective Practice | Teaching Guides | Teaching Commons | DePaul University, Chicago

Vanderbilt University

Teaching Portfolios | Center for Teaching | Vanderbilt University

Colorado State University (University of Dayton Teaching Portfolio Guidelines)

University-of-Dayton-Teaching-Portfolio-Guide.pdf (colostate.edu)