

Ad Hoc Committee on Student Evaluations of Teaching

Final Report

Membership

Steve Abraham	HRM, UUP Chapter President
Isabelle Bichindaritz	CSC
David Crider (Co-Chair)	COM
Anne Fairbrother	C&I
Adam Fay	PSY
Sarah Hanusch	MAT
Jennifer Kagan (Co-Chair)	C&I
Greg Ketcham	Assistant Dean, Extended Learning
Mary McGowan	AFL
Sean Moriarty	Chief Technology Officer
Mihai Paraschiv	ECO

Overview and charge

The COVID-19 pandemic overlaps significant technical changes in campus-wide SET data collection and processing. Campus Technology Services (CTS) discontinued processing of Scantron “bubble sheets” in Fall 2019. The rapid shift to remote teaching beginning in Spring 2020 necessitated a shift to online SET deployment and data collection. Without deep investigation, the number of new electronic surveys during this period is unknown, but there are currently a total of 117 unique student response instruments in Blackboard. It is also likely that Google Sheets was also used as an alternative, at the course and/or department level.

The college plans to implement student opinion surveys using AEFIS; in alignment with this effort the Faculty Assembly charged this Student Evaluation of Teaching (SET) workgroup with the following:

1. Gather copies of all department/program SETs;
2. Explore the use of a subset of common questions in all department evaluations (while still allowing for customization);
3. Develop recommendations about moving to an online-only SET system deployed at an institution level, addressing concerns about response rate; and
4. Develop recommendations with respect to personnel review on the use of SETs as part of a larger teaching portfolio in evaluating teaching effectiveness.

1. Summary of Existing SETs

Committee members requested existing survey instruments from department chairs; a total of 38 surveys were collected from 32 departments, including 3 instruments administered by the Division of

Extended Learning in online course sections. Common themes and topic areas were identified in each school, and summarized below.

1.1 Summary of SETs used in the College of Liberal Arts and Sciences

The subcommittee in charge of reviewing the surveys used by CLAS departments noted that although the different student opinion instruments used were all different, some of the most common questions/items that our subcommittee found run along the lines of instructor/course evaluations.

- The instructor is organized and well-prepared for class.
- The instructor successfully communicates subject matter.
- The instructor created an atmosphere in which students felt free to ask questions.
- The instructor responds effectively to student questions.
- The instructor seemed willing to help students outside of class.*
- The criteria on which exams/assignments were graded were clear.*
- The objectives of the course, the course expectations, and the methods are clearly explained.*
- The assigned readings, assignments/classroom sessions contributed to my understanding of the subject.*
- The instructor is willingly helpful when students have difficulty.

In other cases, common questions/items may include overall instructor/course evaluation (e.g., The instructor was effective. Rate the instructor effectiveness. Rate the overall quality of instruction.). But what does "effectiveness" mean? What about "quality"? If these kinds of questions/items are what is being put forth as a common set, we, as a committee, need to clearly point out that such questions/items should be complemented by questions/items about challenge (which appears often), expected grade, elective/required, time spent on work with this class, etc. The reasoning is simple -- a challenged student may be more likely to provide lower ratings and a student taking the course as an elective may be more likely to provide higher ratings.

At the same time, we may want to make recommendations about how questions/items that are vague (i.e., about effectiveness) may be reworded into something that students can clearly measure (e.g., The instructor responds to emails within 24 hours, excluding Saturday and Sunday; Examinations are returned within a reasonable time (i.e., one week) considering the nature of the exam (objective, essay, number of students, etc.); The assigned readings, assignments/classroom sessions contributed to my understanding of the subject; The objectives of the course, the course expectations, and the methods are clearly explained.)

Our recommendations follow in sections 2.3, 3, and 4.

1.2 Summary of SETs used in the School of Communication Media and the Arts and School of Business

The other subcommittee tasked with analyzing SETs examined SCMA and the School of Business surveys. We analyzed the evaluation instrument used by all School of Business departments, as well as those used by the Art, Communication Studies, Music, and Theater departments in SCMA. We

determined that a mix of “recommended” and “best-practice” questions would be ideal for forming a core that could be used across departments. These questions included:

- The instructor clearly communicated the grading system used in the course.
- The instructor presented/explained material in a clear and organized fashion.
- The instructor was available outside of class time or during office hours when students had difficulty.
- The reading/writing/speaking assignments in the course clearly related back to the learning objectives.
- The instructor demonstrated a thorough knowledge of the subject matter of the course.
- We considered these questions to be objective and largely free from potential biases, either toward the instructor or based on the student’s level of success in the course. We recommend that these same standards be applied to other questions used in SETs whenever possible.

1.3 Summary of SETs used in the School of Education

The subcommittee emphasizes four common questions:

- Does the professor maintain a positive and respectful learning environment?
- Does the professor teach with clear, engaging, relevant pedagogy, and use curricular variety?
- Does the content contain concepts, theories, and research relevant to the field?
- Does the professor give feedback and address strengths and areas for growth?

These common questions cover four broad themes: professor’s professional dispositions, professor’s role in teaching, course’s content and relevance, and assessment and feedback (e.g., professor has clear criteria for evaluation, gives constructive feedback).

Moreover, these themes are specifically tied to the School of Education’s Conceptual Framework, which tracks the extent to which teaching addresses/encourages knowledge formation, practice, reflection, collaboration and leadership, social justice, and authentic learning.

The subcommittee also reports that SETs used across the departments within the School of Education differ to some extent. For example, SETs used by the Counseling & Psychological Services (CPS) department require students to grade both the course and the instructor. The Educational Administration (EAD) SETs focus on leadership roles in schools. Lastly, the SETs used by the Health Promotion & Wellness (HSC) department ask students to evaluate the overall experience in the course, the extent to which they were presented with the opportunity to enhance knowledge, practice, and skills, as well as how valuable the course is to their personal and professional life.

1.4 Summary of SETs used in the Division of Extended Learning (online course surveys)

The Division of Extended Learning has deployed, collected and distributed survey results for all fully asynchronous courses since SUNY Oswego migrated to Blackboard. Prior to that point, surveys administered to online courses were managed by Institutional Research.

The survey asks students to reflect on elements of course design and instructor facilitation. Areas assessed include:

- learning objectives were clearly stated, and related to activities and assignments
- clarity of instructions for assignments
- organization of course materials
- timeliness of instructor grades and feedback on assignments
- timeliness of instructor response to student communication.
- instructor's effective use of the online medium
- instructor's effectiveness in explaining concepts
- instructor's enthusiasm for teaching the course
- availability of instructor for additional assistance

2. Common Questions

2.1 Identification of common questions across departments and schools

In the summaries above, we identified certain questions that would be considered common or “best practice” questions to be used across the campus. Departments should have the freedom to add their own questions to the instrument as needed or desired. The committee agreed that instrument questions (common or otherwise) should be objective in nature, to minimize any effects that may come from such things as a student's anticipated grade in the course or implicit biases about gender, race, or other aspects of identity held by the student. The following sections explain the manner in which our recommendations are grounded in the existing literature regarding student evaluations of teaching.

2.2 Literature

Student evaluations of teaching (SETs) have been used in higher education for more than a century as an inexpensive method of evaluating teaching (Marsh, 1987). However, the ratings provided by these evaluations are often unreliable and invalid measures of teaching effectiveness (Uttl, White & Gonzalez, 2017). This review of the literature is going to elaborate on the issues related to SETs, specifically with sections regarding the purpose, student perceptions, faculty perceptions, measurement and equity biases, and ending with recommendations of best practices. For those who are short on time, we recommend jumping straight to the recommendations section, which summarizes the recommendations explained in the previous sections.

2.2.1 Purpose of SETs

SETs serve several purpose in institutions of higher education, including improving teaching through formative evaluation, informing tenure and promotion decisions through summative evaluation, and demonstrating an institution's accountability to its stakeholders (Kember, Leung & Kwan, 2002; Spooren & Christiaens, 2017; Uttl, White & Gonzalez, 2017). Today, the use of SETs is widespread across colleges and universities in North America, in part, because the surveys used are a low-cost way to give a voice to students about their experiences in the classroom (Bacon, Johnson & Stewart, 2016; Murray, 2005; Uttl, White & Gonzalez, 2017). The use of SETs for making tenure and promotion decisions is the most problematic of these purposes due to the extensive measurement and equity biases, as will be discussed in later sections (Jackson & Jackson, 2015; Kreitzel & Sweet-Cushman, 2021).

2.2.2 Student perceptions of SETs

Student participation is a vital component in the administration of SETs. Luckily research on the students' perceptions of SETs has found that students are willing to participate in the process and generally view the process positively (Kite, Subedi & Bryant-Lees, 2015). In fact, Spooren and Christiaens found that “almost 90% of the respondents (strongly) agreed that SET is necessary to provide accountability for teaching quality” (2017, p. 46), and “students are much more likely to be honest on evaluations if they think that evaluations are effective” (McClain, Gulbis & Hays, 2018, p. 380). However, many students expressed uncertainty about whether their professors actually use the opinions shared (Kember, Leung & Kwan, 2002; Spooren & Christiaens, 2017).

In contrast, students feel ambivalent or negatively towards the use of SETs in summative evaluation. Spooren & Christiaens (2017) claim that students are ambivalent about administration's use of SETs, whereas McClain, et al. claim that “students are less honest if they believe [firing and promotions] is the purpose of evaluations” (2018, p. 380). These seem to imply that we get the most reliable ratings from students when instructor's emphasize the formative aspects of the evaluation, rather than the summative aspects.

Moving the administration of SETs online has implications for student's perceptions of them. First, the non-response rates of online SETs is significantly higher than those administered in-person (Bacon, Johnson & Stewart, 2016), which might indicate that students' view the online SETs as less valued by their professors. Also, students perceive less anonymity with the online evaluations (McClain, Gulbis & Hays, 2018; Layne, Decristoforo & Mcginty, 1999), which may lead to the low response rates and less reliable results.

2.2.3 Faculty perceptions of SETs

Although the original purpose of SETs was to provide formative feedback to faculty about how to improve their courses (Marsh, 1987), research now shows that many faculty do not effectively use the results from the surveys. Moore & Kuol (2005) found that SET results tend to demotivate faculty and faculty will often fixate on minor issues. Some institutions provide professional development to faculty members to help them learn techniques for using the SET results constructively (Salerno, 2019).

One explanation for the challenges that faculty face when reading SET results is that most SETs lack a good theoretical foundation. No broadly accepted theory of good collegiate teaching exists (Jackson & Jackson, 2015; Spooren & Christiaens, 2017), which makes it impossible to create a valid instrument. Thus, many faculty feel that their SET results do not reflect their philosophy of good teaching (Burden, 2008; Burden, 2010). This disconnect means that faculty are more likely to disregard SET results that they disagree with.

2.2.4 Measurement bias in SETs

One of the most significant issues with SETs is that the results are usually plagued with measurement bias. There are three central themes with regards to the source of the bias: the design of the survey instruments, the measurement of irrelevant variables, and the analysis of the survey results.

2.2.4.1 Survey instrument design

At many institutions, the instruments for SETs are not verified for reliability and validity (Jackson & Jackson, 2015). As discussed in the previous subsection, validity is particularly challenging to verify because there is an “absence of a general theory of college teaching upon which SET instruments can be grounded” (Spooren & Christiaens, 2017, p. 43). In addition, students lack the expertise to evaluate teaching quality, so at best these surveys reflect student ratings (Hornstein, 2017; Linse, 2017).

Furthermore, no correlation between SET ratings and student learning has been found (Clayson, 2009; Uttl, White, & Gonzalez, 2017). These findings coupled with the finding that faculty feel that their SET results do not reflect their philosophy of good teaching (Burden, 2008; Burden, 2010), seem to indicate that most SETs are not accurately measuring effective teaching. This result was confirmed by Esary and Valdes (2020) who generated simulated data for teacher quality and SET scores, and found that SET scores are unable to conclude when one faculty member is better than another. Specifically, “over 27% of simulated faculty members at or below the 20th percentile on SETs were actually above the median of instructor quality” (Esary & Valdes, 2020, p. 15). These results indicate that SETs need to be used with great caution when used to assess the quality of teaching for tenure, merit raises and teaching awards.

Another issue with the implementation of SETs is the rates of non-response bias. Nonresponse bias occurs when the respondents to the survey answer differently than those who do not respond, and research indicates that on SETs there are differences in the GPAs and demographics of the respondents and non-respondents (Bacon, Johnson & Stewart, 2016). Response rates for online surveys tend to be much lower than the rates of in-class surveys (Hornstein, 2017). Bacon, Johnson and Stewart (2016) found that low response rates tend to lead to dichotomous results with low variance, whereas higher response rates tend towards the mean with more variance.

2.2.4.2 Measuring irrelevant variables

There is significant evidence that SETs tend to measure numerous factors that are irrelevant to teaching effectiveness (TEIFs), including course characteristics and student characteristics (Kreitzel & Sweet-Cushman, 2021; Uttl, White & Gonzalez, 2017). Many of these characteristics are not within the instructor’s control.

Some of the course characteristics that have been found to have significance on SET results include course meeting time, class size, the workload for the course, and the academic discipline (Kreitzel & Sweet-Cushman, 2021; Spooren & Christiaens, 2017; Uttl & Smibert, 2017; Uttl, White, & Gonzalez, 2017; Uttl, White & Morin, 2013). The academic discipline is the most important characteristic for administrators to heed, as instructors of quantitative courses or natural science courses have statistically significantly lower SET results than those teaching qualitative courses (Kreitzel & Sweet-Cushman, 2021, Uttl & Smibert, 2017; Uttl, White, & Gonzalez, 2017; Uttl, White & Morin, 2013). Additionally, many characteristics of the instructor such as race, gender and accent are also significant, as will be discussed in more detail in the equity bias subsection.

Student characteristics that are significant on SET results include, the students’ prior interest in the course, and the grades in the course (Kreitzel & Sweet-Cushman, 2021; Spooren & Christiaens, 2017; Uttl, White, & Gonzalez, 2017). Students who believe that the SET will help improve their school are

more likely to complete the survey and to be honest about their opinions (Bacon, Johnson & Steward, 2016; McClain, Gulbis & Hays, 2018). Other characteristics that can be significant include whether the instructor brings treats on evaluation day, or the instructions that the instructor gives prior to filling out the results (Farreras & Boyle, 2012; Kreitzel & Sweet-Cushman, 2021).

2.2.4.3 Analysis of results

SET results are often reported as means and standard deviations, sometimes with comparisons to other faculty members. However, SET ratings are not normally distributed, usually with a strong positive skew (Jackson & Jackson, 2015), meaning that the majority of ratings are high, possibly resulting in scores of 4 out of 5 ending up in the lowest quartile of scores (Salerno). Means and standard deviations are impacted by skew, and thus are not appropriate measures for these distributions. Non-parametric statistical tests should be used for comparison, and reporting the median and the mode is more appropriate than the mean (Kreitzel & Sweet-Cushman, 2021).

Furthermore, using SET results to compare faculty members to each other is not a statistically sound practice, and may result in prejudicial results against minority faculty members (Kreitzel & Sweet-Cushman, 2021). As described earlier, SET instruments are often unreliable and measure several extraneous variables, which contribute to the inconsistencies between SET results and actual teaching ability (Esarey & Valdes, 2020). There are processes for normalizing and reducing the bias in the data results to make the results statistically valid (Esarey & Valdes, 2020).

2.2.5 Equity bias in SETs

SET results have been shown to have bias against certain faculty including: women, faculty of color, international faculty, and possibly other factors. The bias is most pronounced in the qualitative comment sections, however, evidence of bias can be observed in the quantitative results as well (Boring, Ottoboni, Stark, 2016; Kreitzel & Sweet-Cushman, 2021).

Gender bias is particularly well studied, with several studies observing differences in SET results along gender lines (Boring, et al, 2016; Kreitzel & Sweet-Cushman, 2021; MacNeill, et al., 2015; Spooren & Christiaens, 2017; Uttl, White, & Gonzalez, 2017). Kreitzel and Sweet-Cushman summarized the literature saying “women and men appear to be evaluated-by students of both genders- through the lens of gender stereotypes” (2021, p. 6). This means that women are rated more highly for exhibiting warmth and sensitivity, rather than teaching prowess. Of particular interest are the studies reported by MacNeill, et al. (2015) and Boring, et al. (2016) who found that in online courses, where the gender of the faculty could be controlled, students gave higher ratings to the instructor identified as male than to the instructor identified as female, and the actual gender of the instructor was not relevant.

SET results also show bias against faculty of color, although this area of research is less well established because of underrepresentation in academia as a whole (Kreitzel & Sweet-Cushman, 2021). Faculty of color are often rated worse than their white colleagues, especially Black professors (Reid, 2010; Smith & Hawkins, 2011). It is theorized that this bias is because “people of color may also be punished more for intersectional stereotype nonconformity” (Kreitzel & Sweet-Cushman, 2021, p. 6). In a related issue, faculty with accents receive lower ratings on SETs (Fan, et al., 2019; Kreitzel & Sweet-Cushman, 2021).

Other forms of equity bias may also be present in SET results, however the research evidence is too scant to be conclusive. One experiment concluded that LGBT faculty with strong teaching were often rated lower than faculty with an unspecified orientation (Ewing, et al., 2003). Other possible biases such as age, disability, pregnancy and motherhood are insufficiently studied in the context of SETs (Kreitzel & Sweet-Cushman, 2021).

2.3 Recommendations from the literature

This section contains the recommendations from the literature gathered into one location. This section will not address why these recommendations are suggested, as those reasons are provided in the discussions above. The recommendations are divided into three subcategories, recommendations for implementation, survey design and faculty evaluation.

2.3.1 Recommendations for Implementation

To improve response rates of online SETs, “schools should adopt consistent policies across all classes with the goal of achieving uniform response rates across all classes” (Bacon, Johnson & Stewart, 2016, p. 102). Specifically, the software used and the timeline for administering the survey should be consistent for all courses (Kreitzel & Sweet-Cushman, 2021).

2.3.2 Recommendations for Survey Design

There are several recommendations for the design of surveys intended to improve validity and reduce the equity bias against instructors. First, Kreitzel & Sweet-Cushman (2021) recommend calling the surveys Student Ratings of Instruction rather than Student Evaluations of Teaching. The questions should be limited to those involving a student’s experience with a course and instructor (Hornstein, 2017), for instance, do not ask “how knowledgeable is your instructor about the content area?” or for assessments of pedagogy. Furthermore, questions asking for an overall evaluation should be omitted, as they have the strongest equity bias (Hornstein, 2017; Stark & Freishtat, 2014).

Several instruments have been tested for sampling validity and item validity, and characterized as being “well designed) by Spooren, Brockx, and Mortelmans (2013).

2.3.3 Recommendations for Faculty Evaluation

When evaluating faculty teaching performance it is important to use multiple measures such as peer evaluation, teaching portfolios and instructor reflections, in addition to SET results (Benton & Young, 2018; Esarey & Valdes, 2020; Kreitzel & Sweet-Cushman, 2021). Using multiple measures of teaching reduces the impact of equity and measurement biases and provides a more robust picture of the instructor’s teaching ability.

Simonson et al. (2021) advocate the use of teaching portfolios carefully crafted by faculty members. This framework can be accompanied by a rubric faculty members can use to guide how they document their teaching process and effectiveness. Numerous authors have emphasized that the use of a portfolio provides a compilation of evidence of teaching effectiveness that a single source of information cannot provide (Angelo and Cross 1993; Barkley and Major 2016; Seldin 2000; Fink 2008; Seldin, Miller, and Seldin 2010; Richmond et al. 2014; Esarey and Valdes 2020) and that these have proved accurate to

document teaching effectiveness (Gibbons et al. 2018; Smith et al. 2014; Seldin 2000; Drinkwater, Matthews, and Seiler 2017). The proposed portfolio content comprises four pillars, and the instructor chooses how they exemplify each of them. The four pillars:

- course design focuses on designing course materials in alignment with course outcomes (syllabus, course assignments, student work samples, course design rationale, etc.);
- scholarly teaching implements evidence-based practices (course activities examples, summative and formative assessments examples, peer teaching feedback, classroom visit feedback, etc.);
- professional development practices reflective teaching and continuous improvement of teaching (reflections on student course evaluations, list of professional development activities, etc.); and
- learner centeredness uses an inclusive, learner-centered approach (syllabus, inclusive teaching practices, course material examples, classroom visit feedback, student surveys, student evaluations, etc.)

(Simonson et al. 2021). The same framework can be used advantageously both for online and face-to-face instruction. These authors also caution that other evaluative methods, such as peer reviews for example, have been shown to be plagued with biases and lack of validity and reliability as well.

Furthermore, the data from the SETs can be displayed and analyzed in manners that allow them to be used more effectively for formative assessment. First, report the median and mode, and provide the distribution of ratings per question rather than reporting the mean and standard deviation (Kreitzel & Sweet-Cushman, 2021). Second, adjust the ratings and scores to remove any systematic non-instructional influences (see Esarey & Valdes, 2020). Third, remove comparisons of faculty to each other such as providing the mean score of the department or the mean score of all instructors teaching a particular course (Jackson & Jackson, 2015), because “comparisons across faculty members further disadvantage already marginalized faculty” (Kreitzel & Sweet-Cushman, 2021, p. 7).

Finally, when reviewing SET data for multiple faculty members, remember that direct comparison of SET results is not statistically appropriate, and instead SET scores should be indicators of exceptional teacher performance (at both extremes) (Esarey & Valdes, 2020; Jackson & Jackson, 2015). One method of comparing instructors ethically using SET scores is that “instructors could be grouped into mid, above and below classifications based on their normalized cumulative measure of effectiveness” (Jackson & Jackson, 2015, p. 171). This method reduces the risk of good teachers being classified as low performing because of the strong positive skew in the data.

3. Recommendations for online-only SET

SUNY Oswego may wish to pilot deployment and data collection using AEFIS before a wide-scale deployment occurs. The pilot phase can test the ease of deployment, communication to students to participate, data collection and analysis, and results distribution.

It should be noted that the college is also transitioning to a new digital learning environment (DLE), Brightspace by Desire2Learn, with full implementation taking place in fall 2022. Methods of identifying

and soliciting student evaluation participation need to be defined and tested prior to a pilot deployment.

Timing of deployment is discussed in section 3.2 below. Analysis of student responses both before and after final examination dates found that there was little difference in response scores for passing students in terms of timing, but students who failed responded with lower ratings post-exam (Arnold, 2009). This supports the suggestions below regarding deployment timing.

3.1 Suggestions on eliciting response rates

Potential best practices were identified via numerous sources: SUNY-wide practice; input from colleagues at other institutions; and recommendations gathered from educational technology communities of practice (EDUCAUSE, OLC, UPCEA). The following are useful guidance that can be applied:

Key principles:

- Surveys should be mobile friendly
- Students should be given class time to complete in seated courses
- Students who are not in class that particular day should be given an opportunity to complete the survey
- Target response rates at other institutions are ~60%
- A personal invitation from the instructor to students in fully online courses has been shown to increase response rates
- Mid-term or periodic checkpoint course surveys inviting student feedback is linked to higher completion rates of end of term surveys

3.2 Deployment suggestions

- Responsibility for the insertion of questions inside the instrument of each set of questions will fall to the following offices
 - College wide - Provost's office
 - School wide - Dean's office
 - Departmental - Chair's office
- For courses twelve weeks or longer:
 - All SET surveys should be deployed in the final two weeks of the semester. The Provost's Office will ensure the timing in the system is correct, and the system will send out spaced out reminders to students who have not completed their surveys.
- For courses that are shorter than twelve weeks, the deployment will be in the final week and extend for one additional week.
- An email to all students should be sent early in the week, and an automated reminder to those who have not completed should be sent out on Thursday afternoon
- Reminders should be put into the digital learning environment(DLE) where possible
- Faculty should encourage students to complete their surveys during class time.

4. Recommendations for the use of SETs in personnel review

To do: develop recommendations with respect to personnel review on the use of SETs as part of a larger teaching portfolio in evaluating teaching effectiveness.

4.1 Procedure for Assessing Current Review Process

The committee sent out a blanket request to each department chair, asking them to give us a brief explanation of how SETs are used in personnel reviews within the department. This explanation included whether other measures of teaching effectiveness were also used as part of the review process, and the relative weights of these measures, if known. In all, 13 out of 30 departments responded to our request (for a response rate of 43%), with varying degrees of specificity given in their explanations (see Appendix for the chairs' responses). Given this low response rate and different levels of disclosure, this analysis of departments' specific evaluation processes cannot be considered definitive.

4.2 Analysis of Current Department Review Processes

In general, the chairs stated that SETs were used as part of a larger review process that also incorporates peer evaluations of teaching and reflective self-narratives describing how instructors have considered student feedback in adjusting their performance. Chairs avoided referring to a "weighting" process or any other kind of quantitative comparison; the Chemistry and Psychology departments were noteworthy exceptions in this regard. Some chairs acknowledged using SETs in the process of awarding Discretionary Salary Increases (DSI), while others did not. Many chairs indicated a preference for returning to the previous method - used under the paper-only system - of funneling SET data through administrative assistants to the instructors. There were also concerns about separation of SETs for online courses from those for face-to-face courses, which made it difficult to fully evaluate all of a department's instructors. Finally, some chairs raised concerns about qualitative responses being dominated by students who were dissatisfied with their grades.

4.3 Recommendations for SET Use

It is our recommendation that SETs not be used as the entire basis for assessing teaching, but instead be considered along with other evidence of excellence in teaching. There are a number of different components of effective teaching, including course management, instructional design, and classroom delivery. In addition, while SETs may be used as evidence of effective classroom delivery, the literature shows that SETs often lack validity and are biased against certain demographic groups. Therefore, while they are an important part of teaching effectiveness, even information from SETs which are designed effectively should be supplemented with other indicators of teaching effectiveness.

We also recommend that the same AEFIS system be used for each department's SET, so there is uniformity across the campus. In addition, SETs should not be used to compare individual professors against each other, as noted above in Section 2.3.3. Finally, the release of SET data should flow from AEFIS to department offices to instructors in a two-step process, allowing for central management of the data.

References

- Angelo, T. A. & Cross, K. P. (1993). *Classroom Assessment Techniques: A Handbook for College Teachers*. 2nd ed. San Francisco, CA: Jossey-Bass Publishers.
- Arnold, I. J. (2009). Do examinations influence student evaluations?. *International Journal of Educational Research*, 48(4), 215-224.
- Bacon, D. R., Johnson, C. J., & Stewart, K. A. (2016). Nonresponse bias in student evaluations of teaching. *Marketing Education Review*, 26, 93-104.
- Barkley, E. F., & Howell Major, C.. (2016) *Learning Assessment Techniques: A Handbook for College Faculty*. Hoboken, NJ: John Wiley and Sons.
- Benton, S. L. & Young, S. (2018). Best practices in the evaluation of teaching. IDEA Paper, 69. URL: https://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/IDEA_Paper_69.pdf
- Boring, A., Ottoboni, K., & 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
- Burden, P. (2008). Does the end of semester evaluation forms represent teacher's views of teaching in a tertiary education context in Japan? *Teaching and Teacher Education*, 24, 1463–1475. <http://dx.doi.org/10.1016/j.tate.2007.11.012>.
- Burden, P. (2010). Creating confusion or creative evaluation? The use of student evaluation of teaching surveys in Japanese tertiary education. *Educational Assessment, Evaluation and Accountability*, 22, 97–117. <http://dx.doi.org/10.1007/s11092-010-9093-z>.
- Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, 31, 16-29.
- Drinkwater, M. J., Matthews, K. E., & Seiler, J. (2017) How Is Science Being Taught? Measuring Evidence-Based Teaching Practices across Undergraduate Science Departments. *CBE - Life Sciences Education* 16 (1):1–11. doi: 10.1187/cbe.15-12-0261.
- Esarey, J., & Valdes, N. (2020) Unbiased, Reliable, and Valid Student Evaluations Can Still Be Unfair. *Assessment and Evaluation in Higher Education* 45:1–15. doi: 10.1080/02602938.2020.1724875.
- Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLoS One*, 14, e0209749.
- Farreras, I. G., & Boyle, R. W. (2012). The effect of faculty self-promotion on student evaluations of teaching. *College Student Journal*, 46, 314-322.
- Fink, L. D. (2008) Evaluating Teaching: A New Approach to an Old Problem. *To Improve the Academy* 26 (1): 3–21. doi: 10.1002/j.2334-4822.2008.tb00497.

- Gibbons, R. E., Villafane, S. M., Stains, M., Murphy, K. L., & Raker, J. R. (2018) Beliefs about Learning and Enacted Instructional Practices: An Investigation in Postsecondary Chemistry Education. *Journal of Research in Science Teaching* 55 (8):1111–33. doi: 10.1002/tea.21444.
- Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, 4, 1-8.
<http://dx.doi.org/10.1080/2331186X.2017.1304016>
- Jackson, M. J., & Jackson, W. T. (2015). The misuse of student evaluations of teaching: Implications, suggestions and alternatives. *Academy of Educational Leadership Journal*, 19, 165-173.
- Kember, D., Leung, D., & Kwan, K. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment and Evaluation in Higher Education*, 27, 411–425.
<http://dx.doi.org/10.1080/026029302200009294>.
- Kite, M. E., Subedi, P. C., & Bryant-Lees, K. B. (2015). Students' perceptions of the teaching evaluation process. *Teaching of Psychology*, 42, 307–314.
- Kreitzel, R. J., & Sweet-Cushman, J. (2021). Evaluating student evaluations of teaching: a review of measurement and equity bias in SETs and recommendations for ethical reform. *Journal of Academic Ethics*.
- Layne, B., Decristoforo, J., & Mcginty, D. (1999). Electronic versus Traditional Student Ratings of Instruction. *Research in Higher Education*, 40, 221–232. doi:10.1023/A:1018738731032.
- Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation*, 54, 94-106.
- MacNell, L., Driscoll, A., & Hunt, A.N. (2015). What's in a name? Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40, 291–303.
- Marsh, H. W. (1987). Student's evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11, 253-388. [http://dx.doi.org/10.1016/0883-0355\(87\)90001-2](http://dx.doi.org/10.1016/0883-0355(87)90001-2).
- Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry, & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence based perspective* (pp. 319–383). New York: Springer.
- McClain, L., Gulbis, A., & Hays, D. (2018). Honesty on student evaluations of teaching: effectiveness, purpose, and timing matter! *Assessment & Evaluation in Higher Education*, 43, 369-385.
- Moore, S., & Kuol, N. (2005). A punitive bureaucratic tool or a valuable resource? Using student evaluations to enhance your teaching. In G. O'Neill, S. Moore, & B. McMullin (Eds.), *Emerging issues in the practice of university learning and teaching all Ireland society for higher education (AISHE) readings part 3: Developing and growing as a university teaching* (pp. 141-146). Dublin: University of Limerick.

- Murray, H. G. (2005). Student evaluation of teaching: Has it made a difference? Presented at the annual meeting of the society for teaching and learning in higher education. Retrieved from: <https://www.stlhe.ca/wp-content/uploads/2011/07/Student-Evaluation-of-Teaching1.pdf>.
- Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education*, 3, 137-152.
- Richmond, A. S., Boysen, G. A., Gurung, R. A. R., Tazeau, Y. N., Meyers, S. A., & Sciutto, M. J.. (2014) Aspirational Model Teaching Criteria for Psychology. *Teaching of Psychology* 41 (4):281–95. doi: 10.1177/0098628314549699.
- Salerno, A. (2019, May 31). SET Theory: On reading student evaluations of teaching. AMS Inclusion-Exclusion Blog <https://blogs.ams.org/inclusionexclusion/2019/05/31/set-theory-on-reading-student-evaluations-of-teaching/>
- Seldin, P. (2000) Teaching Portfolios: A Positive Appraisal. *Academe* 86 (1):36–44. doi: 10.2307/40252334.
- Seldin, P., Miller, J. E., & Seldin, C. A. (2010) *The Teaching Portfolio: A Practical Guide to Improved Performance and Promotion/Tenure Decisions*, the Jossey-Bass Higher and Adult Education Series. San Francisco, CA: Jossey-Bass.
- Simonson, S. R., Earl, B., & Frary, M. (2021). Establishing a framework for assessing teaching effectiveness. *College Teaching*, 1-18.
- Smith, B. P., & Hawkins, B. (2011). Examining student evaluations of black college faculty: does race matter? *Journal of Negro Education*, 80, 149-162.
- Smith, M. K., Vinson, E. L., Smith, J. A., Lewin, J. D., & Stetzer, M. R. (2014) A Campus-Wide Study of STEM Courses: New Perspectives on Teaching Practices and Perceptions. *CBE Life Sciences Education* 13 (4):624–35. doi: 10.1187/cbe.14-06- 0108.
- Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: The state of the art. *Review of Educational Research*, 83(4), 598-642.
- Spooren, P., & Christiaens, W. (2017). I liked your course because I believe in (the power of) student evaluations of teaching (SET). Students' perceptions of a teaching evaluation process and their relationship with SET scores. *Students in Educational Evaluation*, 54, 43-49.
- Stark, P. B., & Freisheat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research*. DOI: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1
- Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluations of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, 54, 22-42.

Appendix: Department Chairs' Answers Regarding Personnel Review

Atmospheric & Geological Sciences: The department uses SETs as part of the evaluation of teaching in addition to classroom visits and self-reflections. These are somewhat equally weighted. We release the results to the faculty directly from the admin assistant.

Biological Sciences: The department uses student evaluations of teaching along with peer evaluations, materials from courses taught and developed, and materials related to the mentoring of research students when evaluating teaching effectiveness. The department does not use a model/rubric that assigns weights to any of these factors.

SETs are not used for DSI. Under normal paper-and-pencil SETs, the admin assistant gets the results and distributes them to instructors. Written comments are typed before distributing.

Career & Technical Education: The student evaluation is used in the CTE department review process by reconciling students' comments with the faculty's narratives on instructional effectiveness. There is no quantitative data assigned to the students' evaluation; rather, the qualitative approach is adopted in using it for the review process. However, the CTE department is also wary in using the students' comments because most times satisfied students decline to complete the end-of-course survey, while dissatisfied students are quick to submit their responses that are sometimes circumstantial.

The student evaluations are used in conjunction with faculty's narratives on how he/she has enhanced instructional challenges and responded to the students' comments. Furthermore, the faculty's responses to the students' evaluation are also analyzed and weighted relative to how they are reflected in the faculty's continuing professional development plans.

Chemistry: Faculty compares themselves to departmental average in these two data points + are asked to report longitudinal data to see their progress: They are then compared to departmental average as well as themselves to see the improvement over time or whether faculty made any impactful changes after the last feedback/review given to the faculty.

The departmental average is between 4.1-4.3 out of 5, so in general anyone between 3.8-4.4 is considered within average; anyone with scores between 3-3.7 needs small improvements to move up, and anyone below 3 needs serious improvement.

Other things we look into: 1) Student comments are required to be submitted and they are analyzed for patterns of praises or complaints to back the numerical values; 2) Peer evaluations, every faculty must have 2-3 peer evaluations every year; 3) The number of student complaints about the faculty; these can be changing sections, style of teaching, not answering emails, not showing up at office hrs, etc., and if there is a pattern with a faculty compared to others, this is also taken into account; 4) Improvement of the student teacher evaluation scores when it is low to begin with is an important indication of

improvement; 5) Attendance of CELT workshops and implementing new strategies to improve and documenting these changes; 6) Creation of new elective courses which attract a large number of students and a successful STE score.

Use of SET for DSI is the same as other personnel decisions in chemistry.

Before COVID, it used to be run by the department (admin assistant), and the department kept a copy and distributed the faculty results. With that, we also receive a departmental total and average, which is useful for comparison. If an online version is created, we would like to have a central management (tool), which should result in a departmental average for each question and total before sharing with faculty.

Cinema/Screen Studies: The department asks that faculty summarize any quantitative data from their student evaluations, highlight some common feedback, and then share with the appropriate department committee how they used the feedback (positive or not) to modify their teaching. For this department, the student evaluations help to measure faculty growth. The department also uses peer observations to measure teaching effectiveness.

Communication Studies: SETs are used in combination with peer evaluations, experience with course design, and a written narrative that include how the instructor responds to peer/SET feedback. The greatest weight is on that written narrative. An exception would be if there are routine and uniform complaints seen across all of these measures.

We use a similar process for DSI. In DSI, we look for strengths in all five areas, but more specifically ask that individuals focus very specifically on those areas in which they excelled over the previous year. SETs are one of the measures that are used, along with collegial evaluations (both from the dept chair, members of personnel, and/or those requested by the instructor from others), participation in CELT or COLT activities, the narrative the individual constructs surrounding their pedagogy (classes developed, changes made based on workshops or feedback, etc...), and the types of teaching activities in which one engaged. They are a piece of the puzzle, but a small piece.

Prior to COVID, the teaching evaluations went to the department's administrative assistant. That person would send out the objective feedback via email and then compile the short-answer responses into one document that was shared later. Currently, since teaching evaluations have been put online, they are administered via Blackboard. Once the semester is over and grades are submitted, the results are available to the instructor in Blackboard. This is an imperfect system for several reasons: the instructor alone has access to the SETs, meaning that the chair and/or personnel have to request access. Additionally, we get much lower student participation than when they are offered in-person. Finally, since no one else sees the report, there lacks checks-and-balances. While I don't think that individuals *do* tamper with the data, misreport, or otherwise provide only a piece of the results of the SET, one *could* if one wanted.

Computer Science: Course evaluations are devised primarily as a way for faculty to gather suggestions and feedback to improve their teaching, and are structured and worded in sometimes unconventional ways (sometimes making sense only for CS courses) to improve their effectiveness in doing so. They are used as one component of personnel evaluation (along with peer evaluations, analysis of student performance on assessed learning outcomes, and other materials) only to the extent of checking for exceptional patterns or problems. The department does not in general make fine-grained comparisons of items across instructors or courses, because they are aware of well-known tendencies for responses to vary with course difficulty, audience, instructor gender and other characteristics. However, faculty are free to statistically summarize results in DSI and promotion materials, and often do so.

Criminal Justice: SETs are used in reviews in the pre-tenure period and for promotion requests. They are used for illustrative purposes in making an argument along the teaching effectiveness continuum.

Economics: SETs do play a role in personnel reviews. The department generally gives more weight to colleague observations, but there is no formal weighting process.

Math: Effective teaching should be documented by: Summaries of student evaluations and written comments; Classroom visits by colleagues; Evidence resulting from student activities such as presentations, projects, papers; Success rates of students over time on common finals.

Modern Languages & Literatures: Peer evaluations of classes are used in conjunction with student evaluations for personnel review. The department also examines syllabi, lesson plans, and materials in the faculty members' teaching portfolio, including self-reflection on teaching. Other instruments that may help would be Gen Ed assessment of courses, cultural assessments required for adolescence education majors, and the faculty members' accuracy in assessing the students' writing and speaking skills in a modern language. The department conducts oral and written proficiency assessments of our majors once a year. Same procedure used for DSI.

The results are released to the instructor before the next semester begins and after students have accessed their final grades. Previous to the past two years, our administrative assistant made the paper evaluations and quantitative data available to the instructors after students accessed their final grades.

Psychology: Evaluation will be made using several kinds of information, including: a) evidence of classroom and course effectiveness, b) evidence of course development and development of pedagogy, instructional materials, syllabi, etc., c) evidence of effectiveness in mentoring and supervising students in such academically relevant activities as independent research, research practica, internships, and other activities directly relevant to the total education of our students (e. g., Psychology Club, etc.). No one source of information (e. g., the CRS) should be dominant in the assessment of teaching.

Anonymous Student Ratings of Teaching (the CRS): Faculty members are expected to have CRS scores near the mode of the department, which is typically near the 4-point level on a five-point rating scheme on the instructor summary item of the CRS. Faculty members not reaching this goal will need to describe

the efforts that are being made to improve teaching. Faculty members should consistently have ratings near the mode of the department after having been a member of the department for two or three years and especially in those courses that the faculty member has taught for more than 2 semesters.

We use the same process for DSI, and all faculty give us their SET data. It goes to them first, and then they provide it to the department.

Political Science: *(quoted from their bylaws) The Department uses multiple methods of evaluating teaching, including student evaluations, peer observations, review of syllabi, and self-assessments. Tenure-track faculty should consult regularly with their mentors, as well as senior faculty, regarding the progress of their teaching. In addition to evaluations and observations, criteria for establishing teaching effectiveness can include:*

- *new course design and curriculum development activities;*
- *supervising individual student research, including facilitating student presentations outside the university (documented by review);*
- *innovative pedagogical practices as reflected in syllabi, course assignments, etcetera;*
- *excellence in advising;*
- *critical reflection on teaching;*
- *professional development activities linked to teaching.*

In 2017, our department rewrote our student course evaluation tool, so that we could get more meaningful feedback from students on questions they were able to assess. Until 2020, faculty included a table summarizing their SETs quantitative results by question and by course on their retention and/or promotion documents. We also expected a narrative section in which faculty would critically reflect on their teaching. In this section faculty would often provide examples of the qualitative feedback from SETs. We have one pre-tenure faculty member, and we will not be using the SETs from the first COVID semester (SP 20), or the last two semesters as comparable to pre-pandemic teaching.

Prior to 2020, we used paper evaluations that were processed by Administrative Computing. That office would return a PDF of the quantitative summary to the faculty member directly. The Administrative Assistant would compile a list of the qualitative feedback and return it to the instructor, copying the chair. Later, Administrative Computing would compile a summary list of SET results by course for the department, and the Chair would share that out with the department. Please note this was only the process for courses taught face-to-face. We have a long-term adjunct who teaches an online section of POL 205. Extended Learning has always delivered its own version of an SET, which was not shared automatically with the department chair, and which also makes us unable to compare SETs among our courses. I am against this artificial separation of online and face-to-face courses.

During the pandemic, our department moved to using a Google Form evaluation, and we have not institutionalized a process for sharing out those results. We know that a change is coming with AEFIS delivering SETs. We have some concerns about this.