Report of the Phase 1 Teaching Evaluation Group

Table of Contents

Executive Summary **2**

Introduction **3**

Findings of the Working Group **4**

Concerns with Student Teaching Evaluations 4

Guidelines for the Use of Student Evaluations of Teaching 6

Form & Design of SETs 6

Interpretation of SETs 6

Other Modes of Assessing Teaching 7

Letters and Testimonials from Students 8

Surveys of Students Who Have Worked with Faculty in Capacities Other than Classroom Teaching 8

Oral Testimony / Interviews with Students 8

Public Solicitation of Student Feedback 8

Letters from Colleagues 8

Peer Observation/Peer Evaluation 8

Evaluation of Course Materials / Teaching Portfolio 9

Self-assessment of Teaching 9

Student Self-assessment 9

Combination of Peer Mentor and Peer Observation 9

Observation of a Peer's Teaching by the Faculty Member 9

Tracking of New Methods Tried 9

Research on Other Modes of Assessing Teaching 10

Formative Assessment vs. Summative Evaluation 11

Policies for Summative Evaluation 12

What is Being Assessed 12

Weighting of Teaching 12

Reliance on Student Teaching Evaluations 13

L&C Faculty Survey Results 14

Final Reflections/Recommendations **18**

References **20**

Appendix A: Articles and Bibliographies (Google folder)

Appendix B: Reports and Evaluation Materials (Google folder)

Appendix C: Peer Institution Policies and Practices (Google folder)

Executive Summary

The charge to the faculty working group on the evaluation of teaching, phase one, was to study existing practices for teaching assessment and to explore research findings on their efficacy and suggestions for their use. We summarize empirical research on student evaluations of teaching (SETs) and proposed guidelines for using them. We review other assessment modes and research on their efficacy, and examine policies for assessment modes adopted elsewhere. We present the results of a survey of Lewis & Clark faculty satisfaction with our current practices, and conclude with reflections and recommendations.

Studies of SETs find that they do not accurately measure teaching effectiveness or student learning, but instead measure students' attitudes toward the course and professor. Strongly influenced by students' expectations of what grade they will receive, student evaluations often show evidence of gender, age, and/or racial bias. Many bodies recommend that they be used only in concert with other instruments, and that they be used primarily to track an individual faculty member's record over time, and not to compare faculty, courses, disciplines, etc. Many other modes of assessment exist, including letters from students and/or colleagues, peer observation, and self-assessment. The reliability of these other modes has been less well-studied, with the most research being done on peer observation/evaluation.

It is important to distinguish between assessment used for formative development of teaching and summative evaluation used for personnel decisions. Many institutions use different modes of evaluation for one vs. the other. There are inconsistencies and lack of clarity in whether institutions' policies are intended to assess "teaching excellence," "effectiveness," or "success," as well as whether and how these terms are defined. Institutions vary in how teaching is weighted in personnel reviews relative to scholarship/creative work; many of our peer institutions give teaching the highest priority, rather than equal weight. In addition, Lewis & Clark is out of step with our peers in regarding SETs as the most important source of information about teaching excellence.

Lewis & Clark faculty are not very satisfied with our current SETs; our survey provides baseline information that can be used to assess the effect of any changes we may implement.

We recommend that the phase two group develop a definition of "teaching excellence" that reflects our values as an institution. We also believe it is time to reconsider our uses of formative assessment and summative evaluation, which are currently intermingled. Finally, we recommend that phase two address our over-reliance on SETs. The research that we carried out should provide ample guidance for each of these steps.

Introduction

During the spring semester of 2020, Dean Bruce Suttmeier convened a faculty working group to gather information on the assessment of teaching. This first phase will be followed by a second task force that will consider possible changes to our assessment methods. The phase 1 group comprised faculty from across the college: Lyell Asher (English), Paulette Bierzychudek (Biology), Diana Leonard (Psychology), Magalí Rabasa (World Languages & Literatures), and Tamily Weissman (Biology), as well as Molly Robinson (World Languages & Literatures), Director of the Teaching Excellence Program and Daena Goldsmith (Rhetoric & Media Studies), Associate Dean for Faculty Development.

Our charge was to conduct research on teaching assessment, including but not limited to the following topics: well-regarded and widely used practices for the evaluation of faculty teaching; current practices of peer institutions; current research on faculty development and evaluation practices; alternative student evaluation instruments; and policies and handbook language governing the role of teaching assessment in tenure and post-tenure review.

We quickly discovered that “evaluation of teaching” glosses over an important distinction between a) assessment used for formative development of teaching, and b) evaluation used to make personnel decisions (e.g., tenure and promotion, salary and performance review). In this document, we will use the term “formative assessment” to refer to uses of information about teaching intended to help faculty improve their teaching practices, and “summative evaluation” to mean uses of information that affect decisions related to hiring, contract renewal, tenure, promotion, or salary.

We also discovered a lack of clarity and consistency in what different teaching evaluation methods and policies actually assess, as well as what they claim to assess. Presumably, the purpose of evaluating teaching is to enhance student learning and to reward practices that facilitate learning for all students; however, student learning is seldom what is measured by the most commonly used methods of assessment. We also found variability in whether an institution’s policies focus on “teaching excellence” (as Lewis & Clark does) or on “success,” “effectiveness,” or some other attribute, and in whether and how these attributes are defined. In our research, we cast a broad net to include practices and policies of institutions that assess “success” or “effectiveness” as well as “excellence.” We are mindful that the methods for doing this are nearly always proxies for determining to what degree instructors are engaged in practices that facilitate learning for all of our students.

Because our tenure and promotion processes stipulate that student teaching evaluations are given the greatest weight in evaluating teaching excellence, and because there is an active discussion among scholars, teachers, and administrators about student teaching evaluation, we begin by summarizing empirical research and contemporary conversation surrounding these instruments, and then review guidelines some institutions and organizations have proposed for using student evaluations. We then review other modes of assessment, and the research on their efficacy. Next, we examine the policies and multiple modes of assessment that other institutions have adopted for formative assessment and summative evaluation. We present the results of a survey we conducted of Lewis & Clark faculty satisfaction with our current practices of assessing teaching. We conclude with some broader reflections and recommendations as we pass this report along to a group charged with considering possible changes in our assessment methods. Appendices to this document provide a bibliography of our sources, and the materials we collected from other institutions.

Findings of the Working Group

Concerns with Student Teaching Evaluations

The most recent and comprehensive studies of student evaluations of teaching (SETs) find that these instruments do not accurately measure either teaching effectiveness or student learning. They tend to measure instead students' attitudes toward the course and the professor under review. Though a few studies from the 1980s purport to show slight positive correlations between SET scores and student learning, a recent re-examination (Uttl et al., 2017) of those studies cast doubt on the existence of even modest positive correlations. In fact, two experiments reported by Stark and Freishtat (2014) found a negative correlation between SET scores and teaching effectiveness, where teaching effectiveness was measured both by student performance on standardized tests and by immediately subsequent course work in the subject area.

If SETs do not measure teaching effectiveness or student learning, what do they measure? Boring et al. (2016) found that SET ratings are "influenced more by instructor gender and student grade expectations than by teaching effectiveness." Macnell et al. (2014) found that both male and female students ranked their online instructor lower if they were told the instructor was female. Data from a French University from 2008 and 2013 (cited by Boring et al., 2016) reinforced those findings.

An experiment conducted in 2014 at Georgia Southern University and reported by Joye and Wilson (2015) described similar bias against female professors, a bias amplified by perceptions of the professor's age. Students lectured to by a professor perceived to be a younger female, for example, rated the professor higher for rapport than when they heard the same lecture from a professor perceived to be an older female. However, their retention of the lecture's content—measured by student performance on a quiz—increased when they perceived the lecture to have been given by this same older female who had been ranked lower for rapport.

Racial bias may also undermine the reliability of SET for evaluating teachers, though the evidence here is less clear, perhaps owing to the difficulties of setting up blind, randomized controlled experiments in academic settings. In response to those difficulties, an experiment reported by Basow et al. (2013) used facial animation software to test student responses to black male, black female, white male and white female "professors" whose lectures were controlled for content and delivery. Contrary to expectations, the predominantly white student body rated the African American lecturer more highly, and exhibited no significant gender bias in the ratings of male and female lecturers. But the students with the white lecturer performed better on the course quiz, leading Basow et al. to conclude that the white lecturer was taken more seriously. There is reason to believe, too, that the racial bias found in laboratory experiments, as well as in controlled experiments in non-academic settings, is at play in student evaluations themselves, the partial result of this 2013 study notwithstanding. And, as Professor Deborah Merritt (2008) pointed out, given the extent to which various non-verbal behaviors and characteristics—such as perceived attractiveness, apparent political affiliation, appearing "relaxed," etc—affect student evaluations, ethno-cultural norms coincident with favored behaviors and characteristics will be rewarded irrespective of their relevance to pedagogical skill or student learning.

Finally, according to Berkeley Statistics Professor Phillip Stark, "the strongest predictor of evaluations is grade expectations" (Flaherty, 2019). This conclusion is based on a number of studies, the most recent and comprehensive of which (Braga et al., 2014) found that SET scores are negatively correlated with subsequent student performance, but positively correlated with the students' grades or expected grades in the class. Braga et al. (2014) supports and extends earlier studies (Carrell & West, 2010; Krautmann & Sander, 1999) in concluding that increasing the rigor of a course in ways that measurably increase student learning generally results in lower scores on student evaluations. This effect is most pronounced among students who are the least academically capable, the study found, and least evident among students who are most capable.

In sum, recent studies that have measured student learning and/or have adopted experimental methods to attempt to establish causality, raise serious doubts that SETs measure teaching effectiveness. They also indicate that SETs may be unduly influenced by an instructor's gender, race, or age. Finally, there is evidence that SETs may actually be negatively correlated with student mastery of course material, at least in curricula that have a sequence of courses and standardized methods of assessing a student's performance in subsequent classes. While the conditions that make for a sound experiment (e.g., standardized measurement of outcomes, random assignment to conditions) are necessary for isolating causation, they do not fully represent the variable contexts in which teaching and learning occur. Nonetheless, if we ask what social scientific evidence reveals about what SETs measure and predict, there is cause for concern in using them to assess "teaching excellence."

SETs may still provide useful information about students' perceptions of and satisfaction with their experience. As we summarize in a subsequent section, most institutions continue to use them in some form. Consequently, those who have researched student evaluations, scholarly organizations, and some individual institutions have developed guidelines for when to use student evaluations and how to interpret their results.

Guidelines for the Use of Student Evaluations of Teaching

We reviewed the websites of the following institutions and organizations, selected because they featured recent comprehensive revisions of, or recommendations for, the teaching evaluation process: Iowa State University (ISU), University of Washington (UW), Stanford, University of Oregon (UO), American Association of University Professors (AAUP), and the American Sociological Association (ASA). URLs for these websites are included in Appendix A. Here we summarize their guidelines for the form, design and interpretation of SETs.

Form & Design of SETs

Some institutions caution against the use of institution-wide common evaluation forms, emphasizing the benefits of evaluation instruments that are context-specific, customizable, and faculty-created (AAUP; Stanford). This caution is motivated by the recognition of problems comparing SET results across disciplines, across courses and even across instructors. These problems are described in the next section (Interpretation of SETs).

In response to concerns about bias in SETs, ISU and UO suggest avoiding “global” or “high inference” questions, such as “rate the extent to which an instructor is organized.” “Low inference,” more specific questions are thought to be more effective in eliciting meaningful feedback and avoiding biased responses. A “low inference” question might ask students to rate specific aspects of an instructor’s organization, such as the extent to which they explain or present a plan for the next class, or signal the transition from one topic or activity to the next.

The AAUP, Stanford, and UO state that SET questions should emphasize the role and responsibility of the student for their own learning, not just the role of the instructor. For example, the University of Oregon has replaced the former end-of-course SET with a three part Course Survey that comprises a Midway Student Experience Survey, an End-of-course Student Experience Survey, and an Instructor Reflection. The student survey no longer uses a point system, but rather offers three rating categories (Beneficial to my learning/ Neutral/ Needs Improvement) and prompts students to provide comments. To emphasize student responsibility for learning, the survey also includes specific questions about how the student supported their own learning.

Interpretation of SETs

Most guidelines offered about the use of SETs focus on their interpretation, suggesting that while the instruments themselves may have shortcomings, the interpretation of the data they yield is an even greater concern. Nearly all of the websites we reviewed emphasize that SETs should be just one part of a holistic evaluation of teaching that includes other modes of assessment as well, such as those we describe in the next section (ASA; UO; UW; ISU). Iowa State's website suggests that effective interpretation of SET data requires developing a shared understanding among faculty and administrators about the use of SETs for different kinds of assessment. Because of the well-documented issue of bias in SETs, UW's website recommends that all faculty and committee members be trained to recognize and address the presence of bias in SET data. There are many resources available for such training and education.

Guidelines for interpretation all attempt to address the potential for bias, as well as other limitations. ISU, UW,and Stanford all recommend looking for patterns across courses, rather than examining data from individual courses. SET data should be examined for evidence of improvement, rather than to assess isolated performance in a single course or term (ISU). These websites also emphasize the importance of recognizing instructional context, with ISU and UW suggesting that SET data only be compared between similar instructional situations (division, discipline, course level, course type, course size, etc.). The ASA further suggests that SET data not be used to compare faculty members to each other or to department or college averages. The ASA, ISU and UW also recommend considering sample size and response rate when interpreting SET data. All of these guidelines are especially important when using quantitative data, which lend themselves to comparison more readily than qualitative data do.

In sum, the guidelines offered by these institutions and organizations assume that the bias and other issues that make SETs unreliable measures of teaching effectiveness can be addressed and mitigated, at least in part, through: 1) the use of additional modes of assessment in combination with SETs; 2) the development of very specific, "low-inference" questions on SET forms; 3) training to reduce bias in interpreting SET data; and 4) recognizing and considering instructional context. However, while many institutions and organizations offer these sorts of guidelines, they do not include assessments or evidence of their effectiveness in mitigating the shortcomings of SETs.

Other Modes of Assessing Teaching

We researched assessment methods used by the NW5 schools and a longer list of peer institutions identified by Institutional Research. These were: U. Puget Sound, Reed College, Whitman College, Willamette U., Centre College, Denison College, Furman U., Gettysburg College, Kalamazoo College, Kenyon College, Lafayette College, Lawrence U., Sewanee: the U. of the South, Skidmore College, Trinity College, and Wabash College. Methods used by additional institutions were also discussed if a committee member had anecdotal knowledge of those practices. We believe that our research was extensive enough to capture the main practices being used to evaluate teaching nationally at institutions similar to Lewis & Clark. Appendix B includes a spreadsheet with details of assessment practices at peer institutions gleaned from their web pages. It also includes separate documents with more detailed summaries of practices among our NW5 peers; for these institutions, we supplemented web research with follow-up questions to their Associate Deans.

Letters and Testimonials from Students

For summative evaluation, it is common for institutions to solicit letters from students of the faculty member under review. Some include only letters from current students; others also request letters from alumni. Some allow the faculty member to choose the students who provide letters; at others, a Chair, Dean or Provost selects students at random. Practices range from a prescribed number of letters solicited from specific categories of former students to less prescriptive methods.

Surveys of Students Who Have Worked with Faculty in Capacities Other than Classroom Teaching

Less commonly, students who have worked with the faculty member in other ways are asked to fill out survey-like evaluations or write letters to be included in a review file. These are primarily: 1) students who have written a thesis or conducted research under the faculty member's guidance; and 2) academic advisees of the faculty member.

Oral Testimony / Interviews with Students

A few institutions conduct interviews with students to solicit feedback about the faculty member's teaching. These students are suggested by the faculty member; suggested by an administrator; or selected at random from the faculty member's course rosters.

Public Solicitation of Student Feedback

At a few institutions, when a faculty member is being reviewed for tenure or promotion, the college places an advertisement in the school newspaper or by email, to solicit feedback on the faculty member from the student body.

Letters from Colleagues

Another common practice for summative evaluation is the solicitation of letters from a faculty member's colleagues. Most often, such letters are based on the colleague's observations of the faculty member's classes. However, they may also be based on a review of the faculty member's course materials, or their broader contributions to the curriculum. Colleagues are chosen for this task in a variety of ways: some institutions designate teaching mentors or a committee to observe and evaluate the faculty member's teaching; some institutions ask all members of the reviewee's department to provide letters; or the reviewee might solicit letters from colleagues themselves. Colleague letters sometimes also assess research and/or service.

Peer Observation/Peer Evaluation

Peer observations are a very common part of summative evaluation. At some institutions, a prescribed person fulfills this role (the chair of the department or an assigned mentor). The prescribed number of visits varies from one time only to many times over a period of years. We assume that class visits lead to some kind of written report or letter, but this is usually not described in detail on institutional webpages. A few institutions train the observers and use standard procedures or forms.

Evaluation of Course Materials / Teaching Portfolio

Many institutions state that they include course materials (syllabus, assignments, tests, etc.) in summative evaluations, but rarely provide details on how the evaluation is done. Some institutions include teaching materials in the files sent to external reviewers in tenure reviews.

Self-assessment of Teaching

Review files commonly include a statement from the reviewee of their teaching philosophy, and reflections on their teaching. A few institutions ask faculty to assess their own practices by using specific checklists or rubrics. At one institution, faculty assess themselves on the use of specific “best practices,” then invite a peer observer to visit their class and provide feedback. Later, after continuing to teach the peer-observed class, they re-assess themselves using the same checklist. The results confirmed that peer evaluations improved the self-assessed use of practices from the checklist.

Student Self-assessment

Some institutions have students evaluate their own behavior during a course, as well as that of the instructor.

Combination of Peer Mentor and Peer Observation

This system was established for long-term, collaborative work on teaching with an assigned, trained peer mentor. Information gathered by the peer mentor is included in the instructor’s file (but is not the sole method of evaluation). This can help temper and “flesh out” information from SETs.

Observation of a Peer’s Teaching by the Faculty Member

This refers to requiring a reviewee to document engaging in regular observation of a colleague’s teaching. In addition to the observation, regular meetings are held for dialogue around teaching. The reviewee may include in the evaluation file a written journal or reflective statement about the observation and dialogue.

Tracking of New Methods Tried

This method involves tracking the active learning techniques a faculty member tries. After every course ends (i.e., once per semester), the faculty member writes a brief reflection on what went well in class, what didn’t, and what they want to try next time. These notes are read by a chair or other supervisor, for purposes of formative dialogue -- not summative assessment.

Research on Other Modes of Assessing Teaching

In contrast to the extensive body of research on SETs, research on the effectiveness of other modes of assessment is sparse. One of the challenges in this research area is determining what would constitute evidence that an assessment or development tool was effective, without an agreed-upon standard of comparison. For example, the Department of Human Physiology at the University of Oregon implemented peer-evaluation; they demonstrated that the use of this tool precipitated increases in scores on the Teaching Practices Inventory (TPI; Dawson & Hawker, 2019). TPI scores represent a faculty member's self-reported activities in the classroom, taken from a list of evidence-based practices intended to encourage learning (Wieman and Gilbert, 2014). The TPI improvements reported by Dawson and Hawker (2019) are promising evidence in support of peer-evaluation as a tool for formative assessment. However, this validation approach is also limited, due to the small scale of the studies and the lack of clear validity of the TPI measure itself.

At first blush, the Teaching Practices Inventory appears to be an objective measure of teaching excellence in science and mathematics. The practices are organized into several categories, including a) transparency of course information (e.g., providing clear learning objectives), b) supporting material provided to students (e.g., practice exams), and c) in-class activities (e.g., pausing to ask for questions; Dawson & Hawker, 2019). However, a one-size-fits-all scoring system loses credibility in the face of the variety of teaching settings to which it would be applied. For example, teachers receive one point for making Powerpoint slides or lecture notes available; however, there is debate on whether this practice supports or hinders student learning (Frey and Birnbaum, 2002; Sidman and Jones, 2007).

A similar tool, the Teacher Behavior Checklist, was subjected to factor analysis by Keeley et al. (2006), who showed that the checklist reflects the typical competence/warmth scales that are known to be influenced by group stereotypes (e.g., women are rated higher on warmth and lower on competence than their male counterparts; Cuddy et al., 2008). Therefore, its use as a student-completed assessment tool may suffer from the same biases as traditional SETs do. However, when self-administered, behavior checklists may provide a standardized way to report and reflect on the practices teachers use in their courses.

Likewise, peer observation/evaluation is sometimes plagued by middling interrater reliability scores (Centra, 1974) and concerns over bias. Lee et al. (2013) discuss the issues with peer review in academia more broadly, including evaluation of research manuscripts and classroom teaching. They conclude that peer review suffers from several forms of bias, including "luck of the draw" in getting a friendly reviewer, and bias pertaining to characteristics and social categories of the reviewee. Further, when used for summative evaluation, peer observation of teaching can have unintended negative consequences. For example, its use in promotion and tenure review may encourage a competitive spirit among faculty, prompting anxiety over the role of personality and popularity in career outcomes (Brown & Ward-Griffin, 1994).

However, positive benefits for formative assessment have also been investigated. Faculty who have undergone peer evaluation report increased confidence (Carroll & O'Loughlin, 2014) and greater critical self-reflection (Hammersley-Fletcher & Orsmond, 2005). Adequate training of peer evaluators may be a linchpin in whether this form of teaching assessment achieves these benefits, however. Kohut et al. (2007) observed that faculty participants in a multi-college survey were ambivalent about whether their training had been adequate and only moderately trended toward agreement that peer observation reports were valuable/useful. They also tended to *dis*agree that such reports were reliable. In contrast to these concerns, however, at least one pilot test showed strong reliability across class sessions and raters alike when implementing a standardized checklist (Brent & Felder, 2004).

Finally, whereas some faculty report fearing peer evaluation because it breaches the privacy of classroom teaching and opens them up to judgment (Carroll & O'Loughlin, 2014), when executed well, peer evaluation can actually increase collegiality among faculty (Hammersley-Fletcher & Orsmond, 2005; Woodman & Parapilly, 2015). For example, Boye & Meixner (2011) found that participants' self-reported feelings of engagement in their faculty community nearly doubled over a one year period after expanding their peer evaluation procedure from dyads to groups. The process of peer evaluation also seems to encourage new participants to intentionally foster peer mentorship in the future, according to a qualitative review of five faculty dyads undergoing peer evaluation for the first time (Carroll & O'Loughlin, 2014).

Formative Assessment vs. Summative Evaluation

In this section, we describe how different institutions use multiple methods of assessment for formative assessment vs. summative evaluation, and focus in particular on their summative evaluation processes (i.e., for tenure, promotion, and/or salary review). We concentrated our research on our NW5 peers and on Institutional Research's official list of peer institutions, but also included other colleges or universities whose practices we found innovative or aspirational. For more detail, see Appendices B and C.

Some institutions use multiple modes of assessment, sometimes selecting particular modes for formative assessment and others for summative evaluation. Midterm SETs are one commonly-used type of formative assessment; at Lawrence University, even end-of-semester SETs are only used for formative assessment. Peer observation is commonly used for both formative assessment and summative evaluation, though its style differs for these two uses. For example, a task force at Rochester Institute of Technology recommended that peer observation for formative assessment be conducted by peers from outside a faculty member's department (whose lack of expert knowledge might be similar to that of students), whereas peer observation for summative evaluation be conducted by departmental peers, both for confidentiality and because of their expertise in the field. Willamette University has "teaching triangles" in which three or four faculty engage in ongoing observation of one another's courses for formative development; these observations are distinct from the peer observations used for summative evaluation (though a reviewee may ask triangle peers to write an evaluation letter).

Some teaching and learning centers offer recommendations about how to combine multiple modes of summative evaluation. They recommend that:

- Summative evaluation should be holistic, combining contextualized interpretation of SETs with other assessment modes. SETs should be considered as suggestive, but not definitive (e.g., ASA, UO, UW, ISU).
- Because SETs are repeated over time, a faculty member's responsiveness to this feedback should be a part of summative evaluation (Stanford, UW).
- If peer observation is used for summative evaluation, it should be standardized across the institution, with training of observers and sharing of common materials and procedures. However, it is important that there be input from faculty into these standards, as well as room for disciplinary flexibility (ASA, RIT).

In summary, institutions typically use multiple methods for evaluating teaching, and distinguish between modes used for formative assessment and summative evaluation. Although Lewis & Clark allows candidates for review and promotion to include other kinds of information in addition to SETs, we have not created standards for the use of these different modes, except to say that SETs are weighted most heavily in tenure and promotion reviews. In addition to determining what modes of assessment we wish to use for formative assessment and summative evaluation, we may also wish to consider developing such standards.

Policies for Summative Evaluation

What is Being Assessed

Some institutions aim to assess "teaching excellence;" others focus upon "teaching success" or "teaching effectiveness." Many institutions define what they believe constitutes excellence, success, or effectiveness (e.g., student learning, faculty motivation, inclusive practices, expertise). For example, in 2017 the University of Oregon undertook high profile changes intended to increase transparency in its teaching evaluation system by aligning its methods directly with this definition: teaching excellence is inclusive, engaged, and research-informed (https://tep.uoregon.edu/teaching-excellence). All aspects of assessment are designed to directly examine the three elements of that definition, in an effort to "align practice with policy." Lewis & Clark's teaching evaluation policies aim to assess "teaching excellence," but we have no definition of what we mean by excellence. We believe it would be worthwhile to consider how we as an institution wish to define excellent teaching, and thus what we wish to assess.

Weighting of Teaching

There is variation among our peers in whether teaching and scholarship/creative work are given equal importance in personnel reviews (as they are at Lewis & Clark) or whether teaching is given greater emphasis. For example, Reed College stipulates that criteria for appointment and advancement are “listed in rank order of importance,” with teaching listed first. Willamette University’s requirements for retention, promotion, tenure, and step increases state that while faculty are evaluated on teaching, scholarship/creative activity, and service, “effective teaching is of paramount importance.” Gettysburg College regards teaching as most important, and Kalamazoo College regards advising and teaching, together, as most important.

Reliance on Student Teaching Evaluations

Our peers vary in how they use SETs and other forms of student input in summative evaluation. Most include SETs as just one kind of assessment. Several of our peers have re-labeled these as “student opinion of instruction surveys” (Furman U.) or “surveys of student judgment” (Reed College). A few of our peers have moved away from using SETs entirely. At Wabash College, SET use is optional for summative evaluation; those faculty who choose to include them in review materials may use forms they have designed themselves. Wabash College incorporates student input through “systematic gathering of student comments.” Lawrence U. makes clear that SETs are used only for formative assessment. Frankly, Lewis and Clark is out of step with our peers in stipulating that “student evaluations are the most important source of information” for assessing teaching, and in requiring that faculty submit for evaluation both numerical data and student comments from all of their courses.

Among institutions that require some form of student input in their summative evaluation processes, we note the following variations among policies:

- Some institutions differentiate between the uses of numerical summaries of student ratings and student comments. For example, at Willamette U., summative evaluation is based only on numerical summaries, though faculty can opt to provide student comments for context. At Reed College, SETs containing student comments are seen only by the instructor; the personnel review committee receives signed student letters that are not seen by the faculty member unless requested.
- Several of our peers give faculty some flexibility in selecting which and how many semesters of course evaluations they submit for summative evaluation. For example, U. of Puget Sound does not require faculty to administer SETs in every class, in every semester. When faculty are reviewed, they must submit SETs from the most recent 4 semesters (for tenure) or the most recent 2 semesters (for promotion or salary review).
- Some institutions allow faculty to supplement standard evaluation questions with items designed to measure outcomes particular to their disciplines, departments, or courses. Kenyon College’s online course evaluations are customizable, with optional items that faculty may add to a core of standard questions. Furman U. requires untenured faculty to use a standard form; all other faculty determine, with their department chairs, whether to use the standard form, a departmental form, or a form they create themselves.
- Several peer institutions solicit letters from students as part of evaluation for tenure, promotion, and/or salary review. These letters may be used instead of SETs or in addition to them.

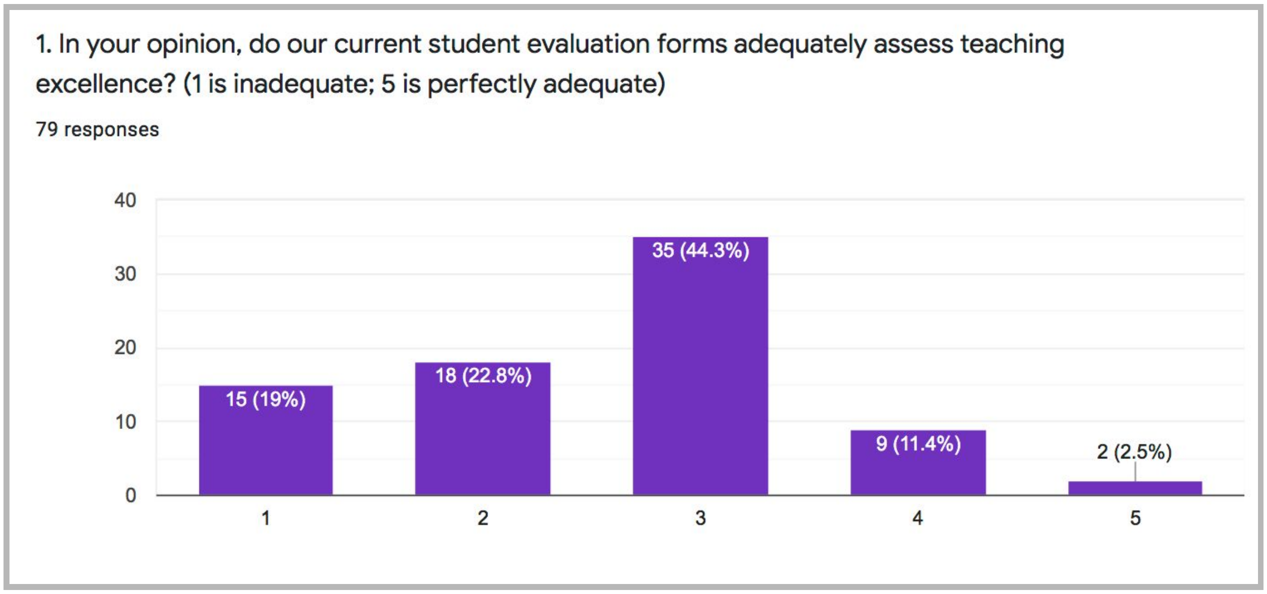
- Many of our peer institutions are also in the process of investigating, proposing, or implementing changes to their SETs and to policies for their use.

L&C Faculty Survey Results

We generated a faculty survey to better understand faculty satisfaction with L&C's current methods of teaching assessment. We asked faculty for their opinions about the current SET forms (questions 1-3), and we probed for their satisfaction with the broader ways of evaluating teaching that are listed in the Faculty Handbook (questions 5-7). Our use of quantitative responses for some of these questions provides a baseline for evaluating the effects of any changes we might make in the future. To learn whether some departments employ additional evaluation tools, we also asked faculty to share any such items (question 4). While our working group also discussed surveying students (for example, asking how well students think evaluations capture their opinions, how comfortable they feel being honest in their responses), we did not administer a student survey. This may be a useful step to consider in the next phase.

The faculty survey was administered via Google Forms, and was open from March 1-9, 2020 (i.e. before campus closed for the COVID-19 pandemic). Eighty-three faculty responded. Results are summarized below (results analyzed at 79 responses).

Question 1 asked faculty "*Do our current student evaluation forms adequately assess teaching excellence?*" The graph that follows shows that most faculty are dissatisfied with our current student evaluations (average = 2.6 out of 1-5 scale). Note that a 5 represents "perfectly adequate," a response that was selected by only 2 of 83 individuals.



1. In your opinion, do our current student evaluation forms adequately assess teaching excellence? (1 is inadequate; 5 is perfectly adequate)

79 responses

Question 2 asked faculty to *"Describe how the student teaching evaluations are useful for assessing teaching*." Many faculty find them useful in some limited ways. The most frequent responses were:

- The written feedback is valuable (21 responses)
- They provide a sense of the student experience, student satisfaction, etc., an important/essential part of evaluation (12 responses)
- They provide insight about broad patterns and general trends (11 responses)
- They allow me to track my trends over time (8 responses)
- Students often provide constructive suggestions that I can implement (7 responses)
- I appreciate student feedback on particular items like my preparation, expectations, feedback, rigor (7 responses)

Question 3 asked faculty to *"Describe how the student teaching evaluations are limited in assessing teaching*." Faculty identified multiple limitations to our current student evaluations. The most frequent responses were:

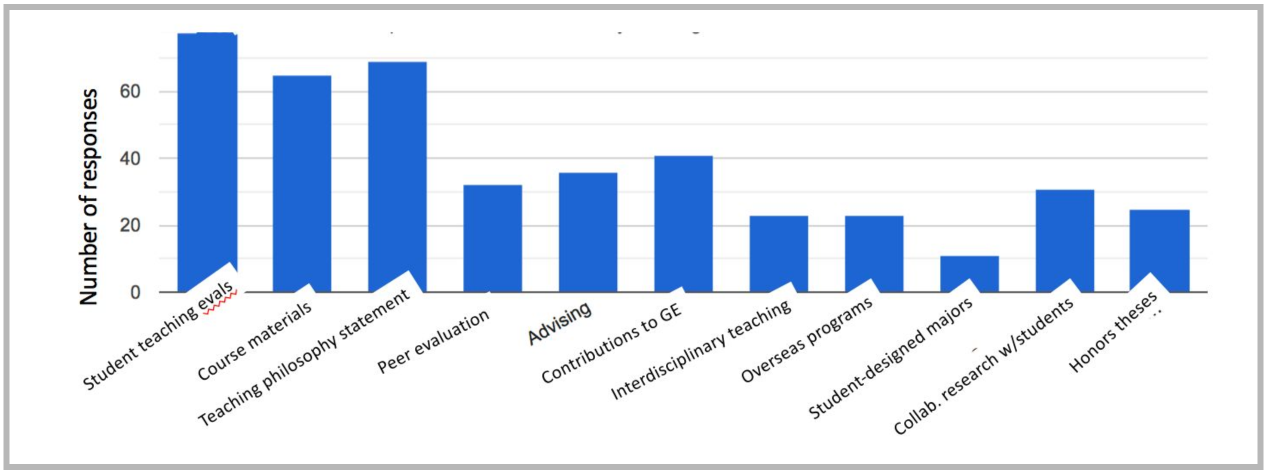
- Scores are unreliable because they are influenced by so many things other than teaching quality: gender and/or underrepresented minority status of the instructor, whether the course is required or elective, introductory or advanced, large or small, its meeting time, its difficulty, etc. (20 responses)
- Evaluations reflect how students feel on one day at the end of the semester when they are tired, stressed, and have not had time to reflect on their learning; students need weeks, months or even years to reflect on what they have learned in a course (15 responses)
- Many students respond emotionally, without reflecting (10 responses)
- Student responses reflect whether a student liked or enjoyed an instructor or course, not whether they learned anything (10 responses)
- Students are not asked to reflect on their own effort, attitude, performance, or whether they ever brought their concerns to the instructor; evaluations seem to shift responsibility for their learning from them to the instructor (8 responses)
- Using evaluations in promotion and review discourages pedagogical innovation and risk-taking; there is a bias against active learning methods ("lazy teaching"); their use in review/promotion tempts faculty to teach to the evaluations; they can be used by department/administration as "weapons" to deny promotion (8 responses)
- Not enough emphasis or nuance on what students have learned (6 responses)
- Students are not skilled at giving constructive feedback; their responses are often too vague to be helpful. Or they might provide diagnosis of a problem, but not how to fix it (6 responses)
- Scores for small classes are unreliable; two sections of the same course can receive very different evaluations depending on mix of student personalities (5 responses)
- Scores are invalid because of how the numbers are analyzed/interpreted (5 responses)

Question 4 asked faculty to list/describe any additional tools that their department uses to evaluate teaching within the department, beyond the formal student evaluations. 24 of 83 faculty reported that their department does not do anything beyond formal student evaluations. Other activities were reported with these frequencies:

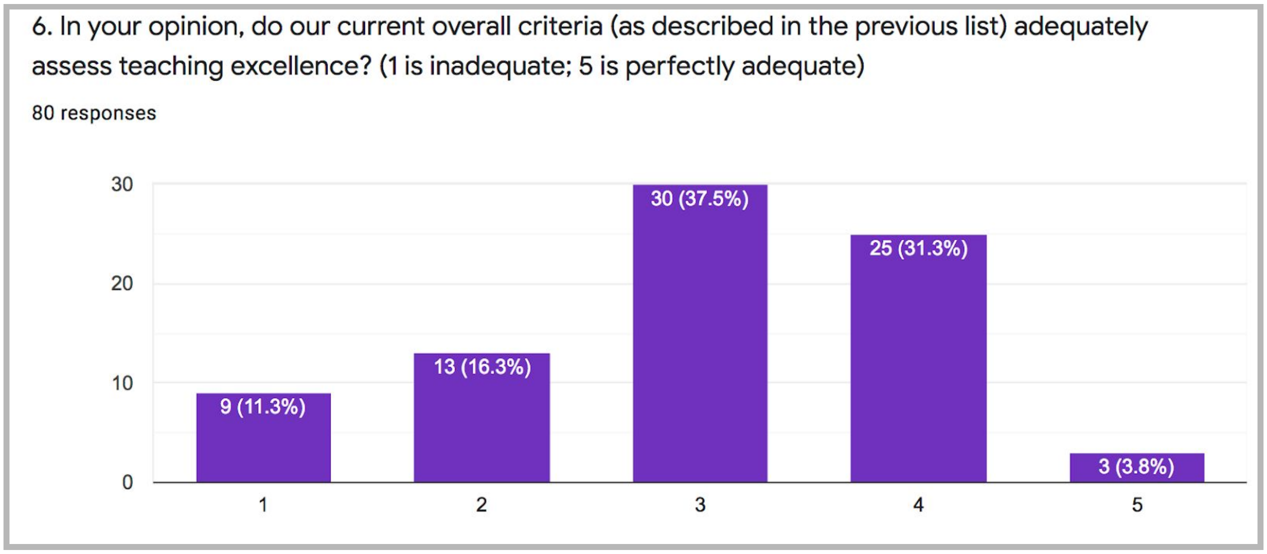
- Peer observation (20 responses -- often reported as sporadic however; no departments reported systematic use except with junior faculty)
- Peer observation of junior faculty members (e.g. for developmental review; 8 responses)
- Use of mid-term evaluations in individual classes (17 reponses) or adding own separate end-of-term questionnaire for students (4)
- Input from course/college alumni (solicited in various ways; 7 responses)
- Informal conversations with students, either during semester or after course ends (6 responses)
- Use of TEP, either through student-partner program, fellow training, or observations - not department-wide however (6 responses). Several individuals mentioned wanting to use this program more.
- Conversations between students and other faculty in department, either informally or as part of an exit interview (4 responses)
- General discussions within the department about pedagogy, syllabi, course outcomes, etc. (3 responses)
- Math and SOAN were reported to use a standardized departmental questionnaire for students in addition to the institution-wide SETs

Methods mentioned by two or fewer respondents included: teaching in each other's classes or team teaching, department mentoring, and standardized national testing. One faculty member has developed a tool in Excel for sorting the data from student evaluations to analyze it more systematically (based on ratings, expected grade in class, major vs. non-major, etc.) and is willing to share this tool with others. Finally, one faculty member mentioned that they use the same questionnaire for students at the beginning of the semester and again at the end, to track change on various measures

Question 5 asked faculty to indicate if they were aware of items listed in the faculty handbook that may be submitted as evidence of teaching excellence. As this graph shows, many faculty are unaware of this diversity of materials (numbers are out of 79 responses).



Question 6 asked for each faculty member's satisfaction with the overall sum of materials that the faculty handbook lists as tools in evaluating teaching (see graph below). There is somewhat greater satisfaction with this overall set of tools (average = 3.0) than for the student evaluations alone (average = 2.6).



6. In your opinion, do our current overall criteria (as described in the previous list) adequately assess teaching excellence? (1 is inadequate; 5 is perfectly adequate)

80 responses

Question 7 asked faculty to *"briefly comment on the list of items we use to assess teaching. What, if anything, do you think is missing? Are there unnecessary items?"* There were many comments suggesting that items in the faculty handbook list are not used consistently to evaluate teaching. There were also questions about how this list was initially generated. A number of individuals suggested that some of the items should not be used evaluate teaching *per se*, and that several of them (for example student-designed majors, interdisciplinary teaching, overseas programs) might fit better in a type of "service to teaching" category, as opposed to direct measures of teaching excellence. In addition, several of the items do not apply to adjunct faculty, such as overseas opportunities, honors thesis mentoring, and in some cases advising.

Other comments addressed the following issues or made suggestions as follows:

- A direct measure of student learning is not included.
- Use of inclusive methods is not assessed.
- Quality of advising is not assessed as much as quantity.
- Students should evaluate themselves as well as the class/instructor.
- Instructors should submit course grades or other measure of course challenge.
- Should assess time spent in office hours, grading, or working with students.
- Should consider enrollment numbers in classes taught.
- Should consider peer consultation outside of classroom, e.g. on syllabus prep.
- Should consider TEP involvement.
- Should separate "service to teaching" from "teaching quality."

- Should measure student success in major or after graduation.
- Should increase alumni involvement - evals of courses/instructors by graduates.
- Should use students in more focused way (e.g. student/teacher partnerships).
- Should include peer observation by an outside expert.

In summary, the faculty survey showed that, in general, there is low satisfaction with SETs (2.6 out of 5). Interestingly, the useful elements of SETs that faculty pointed out in #2 align well with what research suggests SETs can do (e.g. representing students' attitudes toward the course/instructor, longitudinal development of an individual course, self-improvement for individual instructor). There is greater satisfaction with the more complete set of tools listed in the faculty handbook for assessing teaching (3.0 out of 5), but this is coupled with concerns that those measures are not used consistently. Some of the tools listed in the faculty handbook are considered unnecessary and may more appropriately fit in a category of teaching service as opposed to teaching quality. Individuals and departments vary widely regarding the additional tools they use to assess teaching. We anticipate that these survey results will provide a useful baseline against which to .faculty satisfaction following any potential changes to our practices.

Final Reflections/Recommendations

Before we can determine whether our current methods of assessment are achieving our goals, we need to define what we mean by "teaching excellence." What are we attempting to measure with SETs and other materials? What constitutes "excellence"? We presume that a high numerical rating does not *constitute* excellence. What *does* a high SET score reflect or index? These are challenging questions. As we develop the answers, in the next phase of this inquiry, we will need to take care that our definition of excellence is not overly narrow. The purpose of asking these questions is not to suggest that all good teachers look alike or that good teaching is formulaic. However, clear criteria are more transparent, provide a basis for selecting modes of assessment and for creating or refining particular instruments, and can help mitigate bias.

Resolving to define what we mean by excellence requires us to be explicit about our values and provides an opportunity to examine how well our teaching evaluation system aligns with those values. For example:

- Our faculty recently developed a Lewis & Clark Identity Statement that articulates the kinds of opportunities our general education curriculum should provide for all students. These and other statements of our institutional values should inform how we approach the evaluation of teaching.
- If we are committed to creating an environment in which all of our students have the same opportunity to succeed, personally and academically, then our methods of instruction and support must be inclusive. With the support of the Mellon-funded Teaching Excellence Program, Lewis & Clark has committed to inclusive pedagogy--that is, to adopting practices that help all students achieve high standards for learning. Our recent pre-proposal to the Howard Hughes Medical Institute focuses on developing new

ways to assess teaching in the sciences that achieves inclusive excellence. This is a propitious time to re-think how we assess our progress toward this aspiration.

- Relatedly, if we are committed to setting high expectations for student achievement, and then providing instruction and support for our students to meet those expectations, our methods of evaluating teaching should discourage "easy grading."
- Finally, if we value freedom of inquiry, innovation, and creativity our methods of evaluation should not penalize instructors for taking risks or for undertaking change to develop their teaching.

It will also be useful to consider *why* we assess teaching; this may lead us to reconsider our uses of formative assessment and summative evaluation. Our current practices intermingle the two. For example, developmental reviews aim to assist junior faculty improve their teaching, yet are also the bases for contract renewal and salary review. Once tenured, our triennial salary review process is the only required opportunity for tenured faculty to engage in self-reflection about teaching -- presumably this process is both formative and summative. If we assess teaching not only to reward (or penalize) past practice, but also to encourage continued growth, it is useful to consider which kinds of assessment should be used formatively, which should be used summatively, and what relationship these processes should have.

Research on student teaching evaluations and other modes of assessment, examination of practices at other institutions, and results from a survey of our own faculty demonstrate the need to reconsider our over-reliance on SETs as well as the challenges of developing better alternatives. SETs are poor indicators of student learning; even worse, they can introduce inequity into summative evaluations. Well-designed and validated quantitative evaluations can be an efficient source of student feedback, especially when students are asked "low inference" questions about attributes that they can reliably observe. Likewise, there are guidelines for the interpretation of numerical data that are statistically sound (e.g., attend to distributions, sample sizes, and response rates) and fair (i.e. that emphasize tracking an instructor's own scores over time rather than comparing instructors, disciplines or course types). Most institutions continue to incorporate SETs, but as only one mode of assessment.

If in the next phase we decide to change our current methods of assessment, we should do all we can to attend to available research. We should also be prepared to monitor whether and how well any changes are accomplishing our objectives (i.e., assess our assessment). It is challenging to construct reliable survey items; any proposed changes to our current SETs should be appropriately validated. Other forms of teaching evaluation also have strengths and weaknesses, and can be done well or poorly. We acknowledge it would be inconsistent to criticize SETs because the evidence shows a lack of correlation (or negative correlation) with student learning only to embrace other forms of assessment whose connection to student learning has not been well-documented. Faced with this quandary, it is important to give increased attention to appropriately nuanced interpretation of any assessment data. Triangulation among multiple methods is another approach; although this does not provide evidence of a connection to student learning, it does attempt to balance the limits of one mode of assessment with strengths of another.

Finally, we are not the only institution re-examining our teaching assessment practices and policies and we have gathered materials from other institutions in Appendices B and C. At the same time, our methods and policies must reflect our own distinctive context, encompassing the values we share as a community as well as departmental and individual differences. In the conclusion of their 2014 report, the American Association of University Professors "emphasize the primary role of the faculty in teaching evaluation and warn against the encroachment of 'corporate forms of governance' and the growing reliance on numerically based evaluations." Faculty input and ownership will be essential in the next phase in this process.

References

American Association of University Professors (May-June 2016). How do we evaluate teaching? *Academe.* https://www.aaup.org/article/how-do-we-evaluate-teaching#.Xp8j3shKjIV Accessed 4/21/2020.

American Sociological Association (September 2019). *Statement on Student Evaluations of Teaching.* https://www.asanet.org/sites/default/files/asa_statement_on_student_evaluations_of_teaching_feb132020.pdf Accessed 4/21/2020.

Basow, S. A., Codos, S., & Martin, J. L. (2013). The effects of professors' race and gender on student evaluations and performance. *College Student Journal, 47,* 352-363.

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure Teaching Effectiveness. *ScienceOpen Research.* doi: 10.14293/s2199-1006.1.sor-edu.aetbzc.v1

Boye, A. & Meixner, M. (2011). Growing a new generation: Promoting self-reflection through peer observation. In J. E. Miller & J. E. Groccia (Eds.) *To improve the academy,* Vol. 29, (pp. 18-31). San Francisco, CA: Jossey-Bass.

Braga, M., Paccagnella, M., Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41,* 71-88. doi:10.1016/j.econedurev.2014.04.002

Brent, R., & Felder, R. M. (2004). A protocol for peer review of teaching. *Proceedings of the 2004 American Society for Engineering Education Annual Conference & Exposition,* ASEE.

Brown, B., & Ward-Griffin, C. (1994). The use of peer evaluation in promoting nursing faculty teaching effectiveness: A review of the literature. *Nurse Education Today, 14*, 299-305. doi: 10.1016/0260-6917(94)90141-4

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy 118*, 409-432. oi:10.1086/653808

Carroll, C., & O'Loughlin, D. (2014). Peer observations of teaching: Enhancing academic engagement for new participants. *Innovations in Education and Teaching International, 51*, 446-456. doi:10.1080/14703297.2013.778067

Centra, J. A. (1974). Colleagues as raters of classroom instruction. *Journal of Higher Education, 46,* 327-337. doi:10.1080/00221546.1975.11777047

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map.In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology, 40* (pp. 61-149). Elsevier Academic Press. doi:10.1016/S0065-2601(07)00002-0

Dawson, S. M., & Hawker, A. D. (2019). An evidence-based framework for peer review of teaching. A*dvances in Physiology Education, 44*(1), 26-31. doi:10.1152/advan.00088.2019.

Flaherty, C. (2019, December 9). Busting student eval myths? *Inside Higher Ed*. https://www.insidehighered.com/news/2019/12/09/study-attempts-debunk-criticisms-student-evaluations-teaching

Frey, B., & Birnbaum, P. (2002). Learners' perceptions on the use of PowerPoint in lectures. *Computers and Education, 41,* 72-86.

Hammersley-Fletcher, L., & Orsmond, P. (2005). Reflecting on reflective practices within peer observation. *Studies in Higher Education, 30*, 213-24. doi:10.1080/03075070500043358

Iowa State University Center for Excellence in Learning and Teaching (2020). *Student Evaluation of Teaching Guidelines* (website). *https://www.celt.iastate.edu/teaching/assessment-and-evaluation/student-evaluation-of-teaching-set-guidelines-and-recommendations-for-effective-practice/*. Accessed 4/21/2020.

Joye, S. W., & 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. doi: 10.14434/josotl.v15i4.13466

Keeley, J., Smith, D., & Buskist, W. (2006). The Teacher Behaviors Checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology, 33*(2), 84-91. doi:10.1207/s15328023top3302_1.

Kohut, G. F., Burnap, C., & Yon, M. G. (2007). Peer observation of teaching: Perceptions of the observer and the observed. *College Teaching, 55*(1), 19-25. doi:10.3200/CTCH.55.1.19-25

Krautmann, A. C., & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Education Review, 18*(1), 59-63. doi:10.1016/S0272-7757(98)00004-1

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology, 64*(1), 2–17. doi:10.1002/asi.22784

Macnell, L., Driscoll, A., & Hunt, A. N. (2014) What's in a Name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40, 291–303*. doi:10.1007/s10755-014-9313-4

Merritt, D. J. (2008). Bias, the brain, and student evaluations of teaching. *St. John's Law Review, 82*(1), 235-288.

Rochester Institute of Technology Wallace Center (November 2012). *Evaluation of Teaching Effectiveness: Benchmark Report & Recommendations.* *https://www.rit.edu/academicaffairs/facultydevelopment/sites/rit.edu.academicaffairs.facultydevelopment/files/docs/Evaluation_of_Teaching_Effectiveness.pdf* Accessed 4/21/2020.

Sidman, C. L., & Jones, D. (2007). Addressing students' learning styles through skeletal PowerPoint slides: A case study. *Journal of Online Learning and Teaching, 3*, 448-459.

Stanford University Office of the Vice Provost for Teaching and Learning (2020). *Teaching evaluation and student feedback: Principles of evaluation* (website). *https://evals.stanford.edu/principles-evaluation* Accessed 4/21/2020.

Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research*. doi: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

University of Washington Center for Teaching and Learning (2020). *A guide to best practice for evaluating teaching* (website). *https://www.washington.edu/teaching/topics/assessing-and-improving-teaching/evaluation/* Accessed 4/21/2020.

University of Oregon Office of the Provost (2020). *Revising UO's teaching evaluations* (website). https://provost.uoregon.edu/revising-uos-teaching-evaluations. Accessed 4/21/2020.

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. doi:10.1016/j.stueduc.2016.08.007

Wieman, C., & Gilbert, S. (2014). The Teaching Practices Inventory: A new tool for characterizing college and university teaching in mathematics and science. *Life Sciences Education, 13,* 552–569. doi:10.1187/cbe.14-02-0023

Woodman, R. J., & Parapilly, M. B. (2015). The effectiveness of peer review of teaching when performed between early-career academics. *Journal of University Teaching & Learning Practice, 12*(1), 2.