The following Motions and Documents were considered by the GFC Committee on the Learning Environment at its Wednesday, April 26, 2017 meeting:

Agenda Title: **Report of the GFC Committee on Learning Environment on Teaching and Learning and Teaching Evaluation and the Use of the Universal Student Ratings of Instruction (USRI) as an Evaluation Tool**

CARRIED Motion: THAT the GFC Committee on the Learning Environment approve the recommendations and the report on Teaching and Learning and Teaching Evaluation and the Use of the Universal Student Ratings of Instruction (USRI) as an Evaluation Tool, as revised.

Final Item: 4

Recommendations from the GFC Committee on the Learning Environment on Teaching Evaluation and the Use of the Universal Student ratings of Instruction (USRI) as an Evaluation Tool

With General Faculties Council approval, the Committee on the Learning Environment would like to continue our work examining teacher assessment and evaluation. We believe that "Robust supports, tools, and training to assess teaching quality, using qualitative and quantitative criteria that are fair, equitable, and meaningful across disciplines" is an attainable goal towards fulfilling Objective 13 in For the Public Good: "To inspire, model, and support excellence in teaching and learning."

We plan to use the following recommendations in our work plan:

1) Re-examine the overall goals of teaching assessment and evaluation at the U of A ensuring that these goals:

a. Provide the instructor with feedback to improve their teaching (formative assessment)
b. Provide administrators with evidence of effective teaching for merit, promotion and tenure decisions (summative evaluation).

2) Consult with the Faculties and the literature in order to define qualities and measures of effective teaching and ensure that there is a clear link between these qualities and measures.

3) Examine GFC Policy 111. "Teaching and Learning and Teaching Evaluation" and transition this policy to UAPPOL. In the process, we will:

a. Examine how decisions regarding promotion and tenure can be based on multiple indicators of effective teaching, including course based evaluations and more broadly on other teaching related duties.
b. Support consistent interpretation of multiple indicators of effective teaching across the University.
c. Separate instructor feedback for improvement of teaching (formative assessment) and administrative evidence of effective teaching for merit, promotion and tenure decisions (summative evaluation) in both policy and practice.
d. Develop guidelines for the timing, depth and frequency of summative evaluations.

4) Create a suite of assessment and evaluation tools and supports (for both faculty and administrators) with definitions, examples and specific strategies. In developing these resources we will:

a. Investigate methods for instructors to use feedback to improve their teaching and recommend opportunities for teaching development, support and training.
b. Investigate methods and tools to support administrators in using a variety of assessment and evaluation strategies and recommend opportunities for training.

5) Ensure student input is included in teaching evaluation. In our re-examination of the current methods in which student ratings are collected, we will consider:

a. Using student input for both feedback to improve teaching and for feedback in promotion and tenure decisions (formative assessment and summative evaluation), but separating these two purposes in both policy and practice.
b. Examining when student evaluations should not be used by FEC for merit, promotion or tenure decisions.
c. Shifting the emphasis of some of the student rating questions from teacher to student, looking at participation and learning in addition to instruction.
d. Increasing the flexibility of the student rating instrument to apply to multiple teaching contexts (including various class sizes and levels) and unique needs within Faculties.
e. Creating options within the student rating tool that allow the instructor to contextualize their course.
f. Examining qualitative student comments and methods to optimize their use in teaching evaluation.
g. Continued investigations into bias and student ratings.
h. Standardizing methods to optimize response rates and quality of comments with the electronic student ratings.

i. Providing all students (including those with accommodation requirements or those who have withdrawn from a course) with a fair opportunity to provide feedback.

Summary Report of the Evaluation of Teaching at the University of Alberta

Prepared by:

Sarah Forgie, Vice-Provost (Learning Initiatives) and CLE Chair
Norma Nocente, Associate Director, CTL
L. Francisco Vargas M., Senior Research Coordinator, CTL
Anita Parker, Research Assistant, CTL
Carol Brown, Educational Developer, CTL
Rebecca Best-Bertwistle, Research Assistant, CTL

April 2017

Table of Contents

1. Introduction

The University of Alberta is committed to excellence in teaching. Its institutional strategic plan, *For the Public Good,* pledges to "inspire, model, and support excellence in teaching and learning" (University of Alberta, 2016, p. 21). Evaluation of teaching plays an important role in upholding this commitment by shaping the quality of instruction being offered to students. Universal Student Ratings of Instruction (USRI) questionnaires can provide *formative evaluation*, revealing areas of strength or shortcomings related to aspects of teaching, such as planning, organization, communication, and assessment.

Teaching evaluations also affect the careers of instructors at the University of Alberta, since USRI results are used as *summative evaluation* for faculty annual review, as well as tenure and promotion. This dual purpose of USRIs (summative and formative) is often contentious, because of their perceived weight with Faculty Evaluation Committees (FEC). Consequently, in May 2016 the Committee on the Learning Environment (CLE) was tasked by the General Faculties Council (GFC) to report on research into tools for evaluation of teaching by students in university courses. This was to include a critical review of the USRI, as well as an overview of possible multifaceted evaluation methods, ultimately intending to satisfy the University's institutional strategic plan to "provide robust supports, tools, and training to develop and assess teaching quality, using qualitative and quantitative criteria that are fair, equitable, and meaningful across disciplines" (University of Alberta, 2016, p. 21).

CLE approached their investigation with three questions:

1. What does the research have to say about student ratings of instruction?
2. How are the USRIs and other tools used in the evaluation of teaching at the University of Alberta?
3. What are some approaches for multifaceted evaluation of teaching?

The purpose of this report is to address these questions and provide CLE and GFC with information to guide future decisions on the USRI instrument and multifaceted evaluation of teaching at the University of Alberta.

2. Method

Data for this report were obtained from multiple sources. We reviewed 81 articles relating to the three questions above, beginning with literature referenced in the 2009 CLE report *Evaluation of Teaching at the U of A* (Kanuka et al. 2009), which led us to more recent articles (see Appendix A). We researched evaluation processes by other universities, reviewed University of Alberta reports and documents, and conducted interviews with University of Alberta department chairs (see a full report of interviews with department chairs in Appendix B).

2.1. Student Ratings of Instruction

Investigation of question 1, what research has to say about student ratings of instruction, included a review of reports and documents, which provided background information about the history and current status of teaching evaluation at University of Alberta. These included:

- *Report from the sub-committee on evaluation of alternate-delivery courses* (Erkut & Kreber, 2002);

- *Evaluation of teaching at the U of A* (Kanuka, Marentette, Braga, Campbell, Harvey, Holte, Nychka, Precht, Read, Skappak, & Varnhagen, 2009);
- *AASUA position statement on URSIs* (Association of Academic Staff University of Alberta [AASUA], 2012);
- *Report of the GFC Committee on the Learning Environment subcommittee on the status of the USRIs* (Andrews, Chelen, Connor, Kostiuk, Kwong See, & Milner, 2013);
- *Report of the Renaissance Committee* (Cheeseman, MacLaren, Carey, Glanfield, Liu, McFarlane, Cahill, Garneau, Supernant, & Szeman, 2013); and
- *GFC policy manual.* (General Faculties Council, n.d.).

For this report, Test Scoring & Questionnaire Services (TSQS) at University of Alberta conducted descriptive analyses that generated gender-specific USRI scores using data from the academic years 2011/12 to 2015/16. TSQS also participated in an unstructured interview about the validity, reliability, and use of USRIs at the University of Alberta.

2.2. Evaluation of Teaching at University of Alberta

Investigation of question 2, how USRIs and other tools are used at University of Alberta, included short, semi-structured interviews with department chairs (or their equivalents in non-departmental faculties). These interviews were 35-40 minutes, audio recorded, and used an interview protocol pre-approved by CLE with questions about their experiences evaluating teaching (see Appendix C). Interview participants were also given two sample USRI case studies representing real teaching scores and were asked to interpret the scores within the context of their department (see Appendix D). They were asked to reflect on both score sets as if both instructors were teaching different sections of the same course. All potential interview participants were emailed directly with information about the study, including a research letter of invitation, and were encouraged to contact any member of the research team if they had questions or concerns. Data was collected from January to March 2017.

2.3. Multifaceted Evaluation

Information sources for question 3, approaches to multifaceted evaluation, included:

- University of Alberta reports and documents (listed above);
- *Multifaceted summative evaluation of teaching*, a symposium held in May 2015 at Centre of Teaching and Learning (CTL), University of Alberta;
- *University of Alberta peer review of teaching* (Gibson, n.d.); and
- Interviews with department chairs.

3. Findings

3.1. Student Ratings of Instruction

Information from University of Alberta reports and documents

The 2009 CLE report outlined a number of recommendations related to the USRI instrument and to teaching evaluation more generally, as well as GFC policy (Kanuka et al., 2009). This report reviewed literature from up to 2008 and selected 35 articles providing insights

on the following themes: validity; bias; whether students can effectively measure quality teaching; the need for effective tools; correlations between grades and ratings; the impact of evaluation on quality teaching; and the evaluation of faculty for tenure and promotion.

In 2012, the 2009 CLE report was revisited, and the resulting 2013 CLE report, *Report of the GFC Committee on the Learning Environment subcommittee on the status of the USRIs*, put forward four recommendations, including that the purpose of USRIs needs to be clearly identified, and that GFC policy needs updating. It was also suggested that a "working group be struck to determine how to promote consistent interpretation and implementation of policy" (Andrews et al., 2013).

In 2013, the Renaissance Committee, ratified by the AASUA and the Governors of the University of Alberta, addressed aspects of the terms and conditions of work performed at the University of Alberta. Their report detailed a number of concerns and made specific recommendations related to the evaluation of teaching, including USRIs (Cheeseman et al., 2013). The committee recommended that the University of Alberta design a set of questions on the USRI that evaluate the effectiveness of teaching. There is no evidence to indicate that any of the recommendations from the 2009 CLE, 2013 CLE, or 2013 Renaissance Committee reports were pursued. See Appendix E for a table summarizing the positions and recommendations related to USRIs in University of Alberta policy, documents, and reports.

> *"I'm not entirely happy with the question set, I don't think anybody is, it's been forever. I've been on committees for years on this campus and this just keeps coming up so, it's a flawed system so you have to sort of filter it and understand"* (Department Chair).

Review of the literature

In our review of articles referenced in the 2009 CLE report, as well as articles published thereafter, we organized literature relating to student ratings of instruction into two categories – biases and validity (see Appendix A).

Biases. We divided the biases category into sub-categories of gender, instructor characteristics, the correlation between grades and ratings, nonresponse, and non-instructional factors.

- *Gender.* The literature in this category is extensive and conflicted. Numerous articles in this subcategory report gender differences or no differences in student evaluations of teaching. For example, Boring, Ottoboni, and Stark (2016) concluded that student ratings are "biased against female instructors by an amount that is large and statistically significant." On the other hand, Wright and Jenkins-Guarieri (2012) conducted a meta-analysis of 193 studies and concluded that student evaluations appear to be free from gender bias. The University of Alberta TSQS conducted descriptive analyses and the results showed there is no apparent difference between scores for males (N = 18576, Mdn = 4.53) and females (N = 13679, Mdn = 4.57) for statement 211 ("overall the instructor was excellent").
- *Instructor characteristics.* Article findings in this sub-category, seven articles total, were that: instructor personality positively correlates with student evaluations (Clayson, 2013;

Kim & MacCann, 2016); instructor physical attractiveness positively correlates with student evaluations on RateMyProfessor.com (Felton, Mitchell, & Stinson, 2004); instructor age negatively correlates with student evaluations on RateMyProfessor.com (Stonebraker & Stone, 2015) and instructor age impacts negatively on perceptions of teachers and anticipated rapport in the classroom based on photographs (Wilson, Beyer, & Monteiro, 2014); instructor position (limited term lecturer versus full time faculty) does affect student evaluations (Cho & Otani, 2014); and instructor rank (i.e. achievement of tenure) does not affect student evaluations (Cheng, 2015).

> "*But of course you've heard this one before as well, sometimes it's a popularity contest in that you have some individuals who just because of their personality and the way they do things just appeal to the students*" (Department Chair).

- *Correlation between grades and ratings.* Most literature, seven articles in this sub-category, reported that students receiving higher grades tended to provide more favourable evaluations of teaching. Cho, Baek, and Cho (2015) found this to be true in their research study and suggested that it might be a psychological "gift" from the student to the instructor. However, two articles suggested otherwise, such as an analysis of 50,000 courses by Centra (2003) that debunked the correlation between expected grades and student evaluations.
- *Nonresponse.* Nonresponse bias occurs when students choose not to participate in an evaluation of teaching, and the missing data may cause skewed results. Three articles in this sub-category reported that nonresponse bias does influence student evaluations of teaching. For example, Macfadyen, Dawson, Prest, and Gasevic (2016) uncovered that "respondent pools do not fully represent the distribution of students in courses." No articles suggested otherwise.
- *Non-instructional.* Non-instructional bias occurs when circumstances beyond the control of an instructor – such as class type, time, size, and semester – influence student evaluation of teaching. The four articles in this sub-category varied in their investigations and conclusions. For example, Nargundkar and Shrikhande (2014) studied numerous factors and concluded that the combined impact was statistically significant; Reardon, Leierer, and Lee (2014) determined that class schedule does not affect ratings.

> "*...somebody has to teach the broccoli course, right? Not everybody gets to teach dessert, and especially when you get into courses which are, by design or intent or both, more heavily directed towards application, then you are forced to give more critical feedback and that tends to be unpopular*" (Department Chair).

It should be noted that GFC Policy 111.3 (I) also recognizes student bias may impact the evaluation of an instructor.

Validity. Validity refers to the extent that an instrument or procedure measures what it intends to measure, and the extendibility of the results to other situations. Literature within this category equally supports opposing viewpoints as to whether or not student evaluations of teaching are valid measures of teaching quality; whether or not students have the knowledge, skills, or motivation to measure teaching quality. For example, Grammatikopoulos, Linardakis,

Gregoriadis, and Oikonomidis (2015) found an instrument used in the Greek higher education system to be valid, whereas Lama, Arias, Mendoza, and Manahan (2015) stated that students at an Australian university completed surveys without diligence. A meta-analysis by Uttl, White, and Gonzalez (2016) re-analyzed meta-analytic data from Cohen (1981) and concluded that student evaluations of teaching did not indicate teaching quality. Marsh and Roche (1997) found that student evaluations correlated with those of peers and trained evaluators, whereas Uijtdehaage and O'Neal (2015) reported that students mindlessly evaluated a fictitious instructor, even when a photograph was provided. During this project, our research team was not able to find information on the validity of the USRI instrument at the University of Alberta[1].

Related to validity is the impact of student evaluations on teaching quality. In our review of the literature, five articles were divided as to whether or not results from student evaluations had a positive impact on teaching quality. For example, Makondo and Ndebele (2014) reported that lecturers perceive student feedback as valuable for building their teaching skills, yet Stein, Spiller, Harris, Deaker, and Kennedy (2013) argued that evaluation data is not being used effectively for professional development. In a 2011 survey of 564 academic staff at the University of Alberta, 69.2% of respondents agreed that *qualitative comments* on USRIs helped improve the quality of their teaching; 49.5% stated that the USRI's *quantitative scores* were not helpful in this regard (AASUA, 2012).

Information from other universities

The general consensus that student input should be sought related to their experience with course instruction and the learning environment is evident in the practices of institutions other than the University of Alberta. For example, in 2015 Stanford University introduced a new end-of-term course evaluation instrument that included nine required items and additional customizable, open- or closed-ended questions (Stanford University VPTL, n.d.).

Some institutions use multiple instruments to seek insight on students' perceptions of teaching and learning, as well as the broader context of the student experience. For example, both University of Oxford and University of Sydney have recently adopted "The Student Barometer", which includes the learning experience, living experience, support services, and other areas (I-graduate, n.d.). This measure is administered once per year and aims to "track and compare the decision-making, expectations, perceptions and intentions of students from application to graduation" (University of Sydney, 2016a, para. 2). The University of Oxford also employs department-specific evaluation mechanisms, as well as the "National Student Survey" for undergraduate students in the last year of their program (Ipsos MORI, n.d.; University of Oxford, 2015, p. 7).

University of Sydney uses a "Student Experience Survey" for undergraduate students in their first and final year of their program, as well as a mandatory online "Unit of Study Survey (USS)" with eight required items (six quantitative, two open response) and up to four faculty-specific quantitative items and one faculty-specific open response item (University of Sydney, 2016b). Each faculty can also have up to four USS versions to allow customization of

[1] TSQS measures the reliability of the USRI by comparing medians to the previous academic years.

the survey for different contexts (University of Sydney, 2016c). Taken together, the examples provided here highlight that other institutions value student feedback on the teaching and learning environment and are making efforts to update and improve the instruments they utilize to obtain this feedback.

In summary for question 1, what research has to say about student ratings of instruction, we conclude that the topic of survey tools is prevalent the literature, often around the concerns of biases or validity. It is evident that universities globally value student feedback and are working to implement high-quality instruments. University of Alberta reports and documents have historically addressed the USRI, making recommendations for the instrument and University policy; however, there is no indication suggestions made in these reports have had any traction.

3.2. Evaluation of Teaching at University of Alberta

Information from interviews with department chairs

Interview participants from all faculties other than Faculty of Medicine and Dentistry (FOMD) reported using USRIs scores and comments to evaluate teaching; only a portion of FOMD participants reported using this tool. Department chairs revealed that, although they try to consider all the USRI statements, they focus primarily on USRI statement 221 ("overall the instructor was excellent"), and statement 25 ("overall the quality of the course content was excellent") as indicators of effective teaching.

Most participants stated that they approach the interpretation of USRI results with a contextual attitude, indicating that USRIs should be understood in light of instructor characteristics and non-instructional elements.

> "*To be perfectly honest, in the abstract I don't know what I would say. Without knowing the circumstances, if one of those instructors is in her or his first year of teaching, and the other was an experienced professor, I think that interpretation is dramatically different than if they're both experienced professors or if they're both new professors. I can say, if we look at the overall averages they're both scoring in the lower percentile, and that sort of data, but to be perfectly honest that means very little to me because I think that understanding a person's position is crucial to being able to read any of these numbers*" (Department Chair commenting on sample USRIs).

Participants identified several issues with using USRIs exclusively to evaluate teaching, which aligned with our review of the literature, such as biases with gender, instructor characteristics, and non-instructional factors. Most department chairs voiced their need for additional supports to better evaluate teaching. Although some recommended possible alternatives to supplement USRI scores, they still expressed hope that the institution would provide solutions for their concerns.

> "*That question set doesn't serve the diversity and the kind of pedagogy we have now, and really needs fixing. I think there needs to be a conversation about what this is going to look like over time. I also think the University has to take very seriously the concerns that equity seeking groups have about what happens in teaching evaluations. What happens to women? What happens to visible minority? What happens to people that are perceived to have strong accents? And I think there's a huge responsibility on chairs and people on FEC to really be educated in how much you can extrapolate from USRI*" (Department Chairs).

Participants also raised the issue of using USRIs for purposes of tenure and promotion. The 2009 CLE report mentioned this concern, and our review of the literature included seven

articles concerning the use of student surveys for summative purposes, and misinterpretation of their results leading to incorrect conclusions.

In summary for question 2, 'how USRIs and other tools are used at University of Alberta', we conclude that participants from all faculties other than FOMD consistently use USRIs scores and comments to evaluate teaching. Department chairs focus on one or two statements as a barometer of effective teaching, and although most approach interpretation of results with a contextual attitude, they also recognize issues with the USRI that are consistent with our review of the literature, specifically perceived issues of bias, validity, and concerns about potential misinterpretations of student survey results for the summative purposes of tenure and promotion.

3.3. Multifaceted Evaluation

According to Lyde, Grieshaber, & Byrns (2016), a comprehensive system of teaching evaluation is necessary due to the limitations of student surveys and the complex nature of teaching performance. In our review of articles referenced in the 2009 CLE report, as well as more recently, ten articles recognized the need for instruments that are of high psychometric quality, and also that evaluations should include multiple sources of information, such as surveys, peer evaluations, self-evaluations, focus groups, and more.

Reference to multifaceted evaluation is found in University of Alberta documents and reports discussed earlier. The 2009 CLE report commented that an imprecise definition of teaching excellence in section 111.1 of the GFC policy exacerbates the lack of guidance provided to individual faculties for multifaceted evaluation (Kanuka et al., 2009, pp. 21-22). The 2013 CLE report recommended the creation of a resource to guide faculties with "possibilities and/or examples" of multifaceted evaluation (Andrews et al., 2013).

In May 2015, the Centre for Teaching and Learning (CTL) hosted a symposium entitled Multifaceted Summative Evaluation of Teaching, wherein some recommendations for best practice were brought forward. Key points included:

- University of Alberta policy needs to include a clear definition of teaching excellence, including a specific set of criteria of effective teaching that can be used for purposes of evaluation; these criteria should be shared with faculty, instructors and students.
- Both formative and summative evaluation of teaching should be multifaceted, collecting multiple sources of evidence at multiple times annually.
- A multifaceted teaching evaluation plan should be developed to supplement University policy, including definitions, examples, evaluation procedures, and specific strategies for training and support.

Approaches to multifaceted evaluation

The 2013 Renaissance Committee report highlighted the importance of rigorous, multifaceted evaluation, which was described as information "collected through a variety of methods and assessed at multiple points in time" (Cheeseman et al., 2013, p. 7, 69). "The array can include student ratings of courses, a teaching dossier, peer observations, external reviews of content, reflection of the teacher (self-assessment), administrator reviews of content and course observation, review of published work on teaching Scholarship, and evidence supporting

the reputation of the teacher in the field(s) of instruction, within and without the University" (Cheeseman et al., p. 70). See Appendix F for a table summarizing the positions and recommendations related to multifaceted evaluation in University of Alberta policy, documents, and reports.

Peer review of teaching. Gibson (n.d.), author of *University of Alberta Peer Review of Teaching* (an online article provided as a resource for the 2015 CTL symposium), defined peer review of teaching as "informed collegial assessment of faculty teaching for either fostering improvement or making personnel decisions" and stated that both formative and summative methods were required for comprehensive teaching evaluation (para 5). Gibson explained that while quantitative student questionnaires provide information about day-to-day classroom interaction, peer review can broaden this to aspects, such as "course content, academic rigor and appropriateness of objectives and topics;… subject matter expertise; instructional materials and methods; and, assessment and grading" (para 3). Gibson outlined six phases of summative peer review and provided eighteen appendices of practical resources, such as sample observation tools and reports.

Teaching dossiers (portfolios). A teaching dossier serves "to facilitate the presentation of a faculty member's teaching achievements and major strengths for self-assessment and interpretation by others" (Day, Robberecht & Roed, 1996, p. 1). They are a cumulative record of one's teaching activities and often include: "(a) a statement regarding the faculty member's teaching philosophy, goals, and strategies; (b) a description of teaching (planning, preparing, and teaching courses; assessing student learning; and giving feedback); (c) an evaluation of teaching accomplishments; and (d) suggestions regarding possible changes for future teaching" (Day et al.,1996, p. 1). Teaching dossiers require instructors to gather multiple sources of evidence and define the value of their scholarship in teaching (Cheeseman et al., 2013). Related to summative evaluation of teaching, the 2013 Renaissance Committee report recommended that "a teaching dossier, following CTL standards, should be part of all tenure and promotion packages" (Cheeseman et al., 2013, p. 70). A document from the University of Sydney provides a comprehensive list of data sources instructors may include in a dossier.

Interviews with department chairs. Participants indicated having already implemented some approaches for multi-faceted evaluation of teaching. In-class peer observation was the most commonly used additional source of information, followed by annual instructor pedagogical self-reflections. Some departments chairs have also implemented yearly faculty audits, in which a small portion of their professoriate teaching is evaluated in a more comprehensive way, and using a variety of supplementary sources of information. Participants indicated, however, that they mostly obtain these extra resources on a voluntary basis (only when professors agree to provide them), and even when they do obtain these resources, not all of them bring this information to FEC.

> "*I don't think that's very useful by itself, it's incomplete. I'd feel uncomfortable judging somebody's fate just based on that. I'm not saying it's wrong but it's only one piece. It's one piece of understanding, and we take teaching seriously. It's not just a bunch of simple numbers pouring at us. We don't just look at you're above this number or below this number, and we're done. We're looking at you much more carefully than that, but it's a good start*" (Department Chairs).

They voiced their need for additional institutional supports to better evaluate teaching with a multi-faceted approach, and they hope the institution will provide a solution.

In summary for question 3, approaches to multifaceted evaluation, we conclude that: there are numerous potential evaluative methods in addition to student surveys; multifaceted evaluation is encouraged by several University reports and documents and literature in general, as well as mandated by University policy; yet this has not yet translated into its consistent or formal implementation across faculties en masse.

4. Conclusion

The purpose of this report is to support CLE with their investigation into student ratings of instruction, the use of USRIs and other evaluation tools at the University of Alberta, and approaches for multifaceted evaluation of teaching.

Question 1, what does the research have to say about student ratings of instruction?

Research around student ratings of instruction primarily point to concerns about biases and validity of survey tools and results. The perspective that student feedback is valuable to help ensure high-quality teaching environments, yet that survey tools are imperfect and limited for a comprehensive evaluation of teaching, is shared by universities globally.

Question 2, how are the USRIs and other tools used in the evaluation of teaching at the University of Alberta?

Semi-structured interviews with department chairs revealed that USRIs are the primary source of teaching evaluation information for all faculties except FOMD. Specifically, most department chairs indicated that they start with only one or two statements but they do their best to contextualize the numerical results. Some department chairs expressed concerns around biases, validity, and the potential for misinterpretation of USRI results for summative purposes of promotion and tenure decisions.

Question 3, what are some approaches for multifaceted evaluation of teaching?

Multifaceted evaluation is supported by the literature and is also mandated by GFC policy. However, impeding its University-wide adoption and consistency is a lack of support and time for those responsible for conducting such robust, comprehensive evaluations of teaching. Moving forward, systematic and purposeful evaluation of teaching can only materialize if there are realistic and tangible expectations, and supports (documents, workshops, etc.).

5. References

These are the references used in the preparation of this report, not including our review of the literature. For the latter, see Appendix G.

Andrews, N., Chelen, D., Connor, B., Kostiuk, L., Kwong See, S., & Milner, R. (2013, June 5). *Report of the GFC Committee on the Learning Environment subcommittee on the status of the USRIs.* Retrieved from

http://www.governance.ualberta.ca/en/GeneralFacultiesCouncil/CommitteeontheLearningEnvironm/~/media/Governance/Documents/GO05/LEA/13-14/Reports/Item-5-USRI-Subcommittee-Final-Report-June-2013-FINAL.pdf

Association of Academic Staff University of Alberta (AASUA). (2012). *AASUA position statement on URSIs.* Retrieved from http://www.aasua.ca/wp-content/uploads/2014/03/AASUA_Position_Statement_on_USRIs.pdf

Centre for Teaching and Learning. (2015). *Multifaceted summative evaluation of teaching symposium*. Retrieved from https://www.ualberta.ca/centre-for-teaching-and-learning/events/symposium-series/past-symposia/multifaceted-summative-evaluation-teaching

Cheeseman, C., MacLaren, I., Carey, J., Glanfield, F., Liu, L., McFarlane, L., Cahill, J. C., Garneau, T., Supernant, K., & Szeman, I. (2013, December 9). *Report of the Renaissance Committee.* Retrieved from http://www.renaissance.ualberta.ca/

Day, R., Robberecht, P., & Roed, B. (1996). *Teaching dossier: A guide.* Retrieved from: https://d1pbog36rugm0t.cloudfront.net/-/media/ualberta/centre-for-teaching-and-learning/instructional-resources/teaching-dossier/teachingdossierguide-1.pdf

Erkut, E. & Kreber, C. (2002). *Report from the sub-committee on evaluation of alternate-delivery courses: Continuing discussion.* General Faculties Council Teaching and Learning Committee. Retrieved from https://docs.google.com/document/d/1KpLMK5kN4r6Mp_BSEoYiZ1xOrbdI9shhScBDAc2NPdI/edit

General Faculties Council. (n.d.).*GFC policy manual.* Retrieved from http://www.gfcpolicymanual.ualberta.ca/

Gibson, S. (n.d.). *University of Alberta peer review of teaching*. Retrieved from https://www.ualberta.ca/centre-for-teaching-and-learning/events/symposium-series/past-symposia/multifaceted-summative-evaluation-teaching/peer-review-of-teaching

I-graduate. (n.d.). *The student barometer.* Retrieved from https://www.i-graduate.org/services/student-barometer/

Ipsos MORI. (n.d.). *National student survey.* Retrieved from http://www.thestudentsurvey.com/

Kanuka, H., Marentette, P., Braga, J., Campbell, K., Harvey, S., Holte, R., Nychka, J., Precht, D., Read, D., Skappak, C., & Varnhagen, C. (2009, January 9). *Evaluation of teaching at the U of A: Report of the sub-committee of the Committee on the Learning Environment (CLE).* Retrieved from https://www.ualberta.ca/-/media/ualberta/centre-for-teaching-and-learning/symposium/evaluating-teaching-2009/symposiumevaluating-teaching-at-the-u-of-a-taskforce-report.pdf

Lyde, A. R., Grieshaber, D. C., Byrns, G. (2016). Faculty teaching performance: Perceptions of a multi-source method for evaluation (MME). *Journal of the Scholarship of Teaching and Learning, 16*(3), 82-94. doi: 10.14434/josotl.v16i3.18145

Stanford University Vice Provost for Teaching and Learning (VPTL). (n.d.). *Standard Course Evaluation Questions*. Retrieved from: https://vptl.stanford.edu/teaching-learning/teaching-practices/evaluation-feedback/stanfords-new-course-evaluations/standard

University of Alberta. (2016). *For the public good: Institutional strategic plan 2016-2021.* Retrieved from https://www.ualberta.ca/strategic-plan

University of Sydney. (n.d.). Teaching insight: Possible data sources to draw on when providing evidence about your teaching. Retrieved from http://sydney.edu.au/education-portfolio/ei/programs/teaching_insights/pdf/insight7_evidence.pdf

University of Sydney. (2016a). *Student Barometer (SB/IB)*. Retrieved from: http://sydney.edu.au/education-portfolio/ei/studentbarometer/

University of Sydney. (2016b). *Student Experience Survey (SES)*. Retrieved from: http://sydney.edu.au/education-portfolio/ei/ses/

University of Sydney. (2016c). *Unit of Study Survey (USS)*. Retrieved from: http://sydney.edu.au/education-portfolio/ei/USS/default.htm

University of Oxford. (2015). *Procedures for the Annual Monitoring of Courses.* Retrieved from: https://www.admin.ox.ac.uk/edc/qa/pamc/

6. Appendices

Appendix A: Table of Reviewed Literature

This table contains literature referenced in the 2009 CLE report, as well as more recent articles relating to the evaluation of teaching. Due to varied research methodologies, measures, and results, definitive comparisons and conclusions from the literature is not be possible; however, the depth and breadth of the articles can provide a general idea about current academic perspectives. Black font indicates literature cited in the 2009 CLE report; green font indicates more recent articles. Brief summarizing points from each article are provided.

Click on the links to move directly to each bookmarked section. For abridged abstracts, see Appendix H. For a complete reference list, see Appendix G.

Biases
This category is divided into sub-categories of gender, instructor characteristics, correlation between grades and ratings, nonresponse, and non-instructional. Also, an "other" category includes articles that focused on multiple biasing factors, biasing factors that do not fit into any other category, or biases in general.

Biases, Gender. Most literature, seven articles in this sub-category, reported that an instructor's gender does influence student evaluations of teaching; however, two articles suggest otherwise.	
Gender influences student ratings	**Gender does not influence student ratings**
Boring, Ottoboni, & Stark (2016): ratings are biased against female instructors by an amount that is large and statistically significant Gehrt, Louie, & Osland (2015): female students evaluated female lower-ranked faculty most favorably; male students evaluations were more favorable for lower ranked male faculty, but they did not degrade higher ranked female faculty Huebner & Magel (2015): variances of the class average responses between male and female faculty were higher for male faculty Laube, Massoni, Sprague, & Ferber (2007): the inconsistency on the question of whether student evaluations are gendered is itself an artifact of the way that quantitative measures can mask underlying gender bias MacNell, Driscoll, & Hunt (2015): students rate males significantly higher than females Miles & House (2015): lower ratings for female instructors teaching larger required classes Wilson, Beyer, & Monteiro (2014): lower ratings for older instructors, but more so for females than males	Centra & Gaubatz (2000): only small same-gender preferences found, particularly with females Smith, Yoo, Farr, Salmon, & Miller (2007): male and female students rated female instructors more highly; effect was small but significant due to sample size Wright & Jenkins-Guarieri (2012): SETs appear to be valid and free from gender bias

Biases, Instructor characteristics (appearance, personality, age, and/or rank). Article findings in this sub-category, seven articles total, were that: instructor personality positively correlates with student evaluations; instructor physical attractiveness positively correlates with student evaluations; instructor age negatively correlates with student evaluations; instructor rank does affect student evaluations; and instructor rank does not affect student evaluations.	
Instructor characteristics influence student ratings	**Instructor characteristics do not influence student ratings**
Cho & Otani (2014): students give higher ratings for limited-term lecturers versus full-time faculty Clayson (2013): students' first perceptions of an instructor's personality are significantly related to ratings at the end of the semester Felton, Mitchell, & Stinson (2004): students give attractively-rated professors higher quality and easiness scores Kim & MacCann (2016): students' expressed educational satisfaction was related to perceptions of instructor personality Stonebraker & Stone (2015): age has a negative impact on student ratings of faculty members; begins around mid-forties; offset by attractiveness Wilson, Beyer, & Monteiro (2014): lower ratings for older instructors, but more so for females than males	Cheng (2015): tenure does not have a significant impact on student ratings of teaching performance

Biases, Correlation between grades and ratings. Most literature, seven articles in this sub-category, reported that students receiving higher grades tend to provide more favourable evaluations of teaching; however, two articles suggest otherwise.	
There is a correlation between higher grades and higher ratings	**There is not a correlation between higher grades and higher ratings**
Backer (2012): some students punish academics for failing grades with low ratings Blackhart, Peruche, DeWall, & Joiner (2006): higher ratings given to instructors who give higher grades, and also to graduate teaching assistant rank Boring, Ottoboni, & Stark (2016): ratings are more sensitive to students' grade expectations than they are to teaching effectiveness Cho, Baek, & Cho (2015): students with better grades than their expected grades provide a psychological "gift" to their teachers by giving higher ratings Greenwald & Gillmore (1997): the grades-ratings correlation is due to an unwanted influence of instructors' grading leniency; there are 5 theories of the grades-ratings correlation Maurer (2006): cognitive dissonance may be a theory to explain the grades-ratings correlation Miles & House (2015): higher expected grades may lead to higher ratings	Centra (2003): expected grades generally do not affect student evaluations Gump (2007): questions the validity of research done on the leniency hypothesis

Biases, Nonresponse. Nonresponse bias occurs when students choose not to participate in evaluation of teaching, and the missing data may cause skewed results. Three articles in this sub-category reported that nonresponse bias does influence student evaluations of teaching. No articles suggested otherwise.	
Nonresponse bias influences student ratings	**Nonresponse bias does not influence student ratings**
Kuwaiti, AlQuraan, & Subbarayalu (2016): ratings are affected by class size and response rate Macfadyen, Dawson, Prest, & Gasevic (2016): ratings affected by who is completing the surveys Reisenwitz (2015): there are significant differences between those who complete online student evaluations and those who do not	No articles found.

Biases, Non-Instructional. Non-instructional bias occurs when circumstances beyond the control of an instructor, such as class type, time, size, and semester, influence student evaluation of teaching. The four articles in this sub-category varied in their investigations and conclusions.	
Non-instructional factors influence student ratings	**Non-instructional factors do not influence student ratings**
Kuwaiti, AlQuraan, & Subbarayalu (2016): ratings are affected by class size and response rate Nargundkar & Shrikhande (2014): combined impact of all the noninstructional factors studied is statistically significant Royal & Stockdale (2015): students give lower ratings to instructors of quantitative methods subjects	Reardon, Leierer, & Lee (2014): class schedule does not affect ratings

Biases, Other. This sub-category includes literature that focused on multiple biasing factors, biasing factors that do not fit into any other category, or biases in general.	
The factors influence student ratings	
Blackhart, Peruche, DeWall, & Joiner (2006): varying results for investigation if class size, class level, instructor gender, number of publications (faculty instructors), average grade given by the instructor, and instructor rank predicted teaching evaluation ratings Keeley, English, Irons, & Henslee (2013): found halo and ceiling/floor effects to be present and persistent; (Halo effect occurs when a positive rating on one aspect of the SET influences the other aspects. Ceiling and floor effects are issues when the SET instrument scale is limited.) Merritt (2012): covers biases in general, including race minority Pounder (2007): identifies and organizes factors influencing SET scores Zumback & Funke (2014): students' mood affects ratings	

Validity

Literature within this category equally supports opposing viewpoints as to whether or not student evaluations of teaching are valid measures of teaching quality, whether or not students have the knowledge, skills, or motivation to measure teaching quality.

Student Evaluations are (Mostly) Valid Measures of Teaching; Students are able to measure aspects of teaching quality	**Student Evaluations are not/may not be Valid Measures of Teaching; Students may not be able to measure teaching quality**
Al-Eidan, Baig, Magzoub, & Omair (2016): the faculty evaluation tool was found to be reliable, but validity has to be interpreted with caution because of low response Bedggood & Donovan (2012): student satisfaction does not equal teaching quality; both student satisfaction and student learning are relevant measures Chen & Hoshower (2003): student motivation to participate in SET affects ratings Cohen (1981): student ratings are a valid measure of teaching effectiveness; this is the paper included in a **meta-analysis** by Uttl et al. (2016) Dolmans, Janssen-Noordman, & Wolfhagen (2006): students can distinguish excellent and poor teaching quality Ginns, Prosser, & Barrie (2007): the SET tool studied supports quality assurance and improvement processes at the university Grammatikopoulos, Linardakis, Gregoriadis, & Oikonomidis (2015): provides evidence of a valid SET instrument; evaluating test validity is a continuous process, not a one-time event Khong (2014): SET is a valid instrument in evaluating teaching effectiveness	Brown, Wood, Ogden, & Maltby (2014): students' satisfaction rating is context dependent; objective quality and subjective satisfaction are different things and should be assessed accordingly Chonko, Tanner, & Davis (2002): students focus more on qualities that make a course appealing, not learning d'Apollonia & Abrami (1997): student ratings are moderately valid; however, they are affected by administrative, instructor, and course characteristics Dodeen (2013): validity of SET is questionable Grayson (2015): questions student's ability to give accurate ratings Greenwald (1997): student rating measures have validity concerns Lama, Arias, Mendoza, & Manahan (2015): lack of student diligence when rating instructors raises validity concerns Martin, Dennehy, & Morgan (2013): validity of SET is questioned; student focus groups suggested as an alternative Morley (2012): student evaluations in this study were generally unreliable

Validity, *continued*

Student Evaluations are (Mostly) Valid Measures of Teaching; Students are able to measure aspects of teaching quality	**Student Evaluations are not/may not be Valid Measures of Teaching; Students may not be able to measure teaching quality**
Marsh & Roche (1997): evaluations are relatively valid and unaffected by hypothesized biases; student ratings correlate with those of peer evaluators and trained evaluators McKeachie (1997): student ratings are valid but affected by contextual variables such as grading leniency Nargundkar & Shrikhande (2012): an instrument that was validated 20 years ago is still valid Socha (2013): a SET instrument was found to have overall good reliability and validity with relatively few biases Wright & Jenkins-Guarieri (2012): SETs appear to be valid and free from gender bias	Rantanen (2013): reliability of SET is questionable; multiple feedbacks required Spooren, Brockx, & Mortelmans (2013): the utility and validity of SET is questionable Uttl, White, & Gonzalez (2016): SETs do not indicate teaching quality, **meta-analysis** Uijtdehaage & O'Neal (2015): many students rate instructors mindlessly

Impact on Teaching Quality

The five articles in this category are divided as to whether or not results from student evaluations of teaching have a positive impact on teaching quality.

Evaluation results may have an impact on teaching quality	Evaluation results may not have an impact on teaching quality
Curwood, Tomitsch, Thomson, & Hendry (2015): provide an example of support for academics' learning from SETs Makondo & Ndebele (2014): SETs are beneficial for improving teaching quality	Asassfeh, Al-Ebous, Khwaileh, & Al-Zoubi (2014): students' perceptions include lack of impact of evaluations on teaching behaviors Campbell & Bozeman (2008): questions the effect student evaluations have on teaching quality Stein, Spiller, Harris, Deaker, & Kennedy (2013): there are gaps in the way academics engage with student evaluation

Evaluating Faculty for Tenure and Promotion

Literature in this category includes seven more recent articles (2012 onward) that express concern about the use of evaluation results for summative purposes, misinterpretation of results leading to incorrect conclusions.

Support for use of student evaluations for tenure and promotion decisions	Concerns related to the use of student evaluations for tenure and promotion decisions
Fraile & Bosch-Morell (2015): present a reliable approach to SET interpretation	Boysen (2015): faculty and administrators can over-interpret small variations Boysen, Raesly, & Casner (2014): ratings are misinterpreted by faculty and administrators Jackson & Jackson (2015): concerns with use of SETs for summative purposes Jones, Gaffney-Rhys, & Jones (2015): presents issues if decision-makers use SET results summatively Mitry & Smith (2014): conclusions drawn from evaluations may be invalid and harmful Palmer (2012): presents examples of ineffective responses to evaluation results

Multifaceted Evaluation

This category amalgamates the concepts of effective tools and multifaceted evaluations into one theme, since effective tools provide the ingredients for multifaceted evaluations. The ten articles in this category recognize the need for instruments that are of high psychometric quality, and also that evaluations should include multiple sources of information, such as surveys, peer evaluations, self-evaluations, focus groups, and more.

Berk (2013): covers several issues, including multifactorial evaluations

Cox, Peeters, Stanford, & Seifert (2013): a peer assessment instrument was piloted; formative peer assessment seems important

Hughes II & Pate (2013): present a multisource evaluation method

Iqbal (2013): faculty express concerns with peer reviews

Lyde, Grieshaber, & Byrns (2016): a multisource method of evaluating is a useful tool

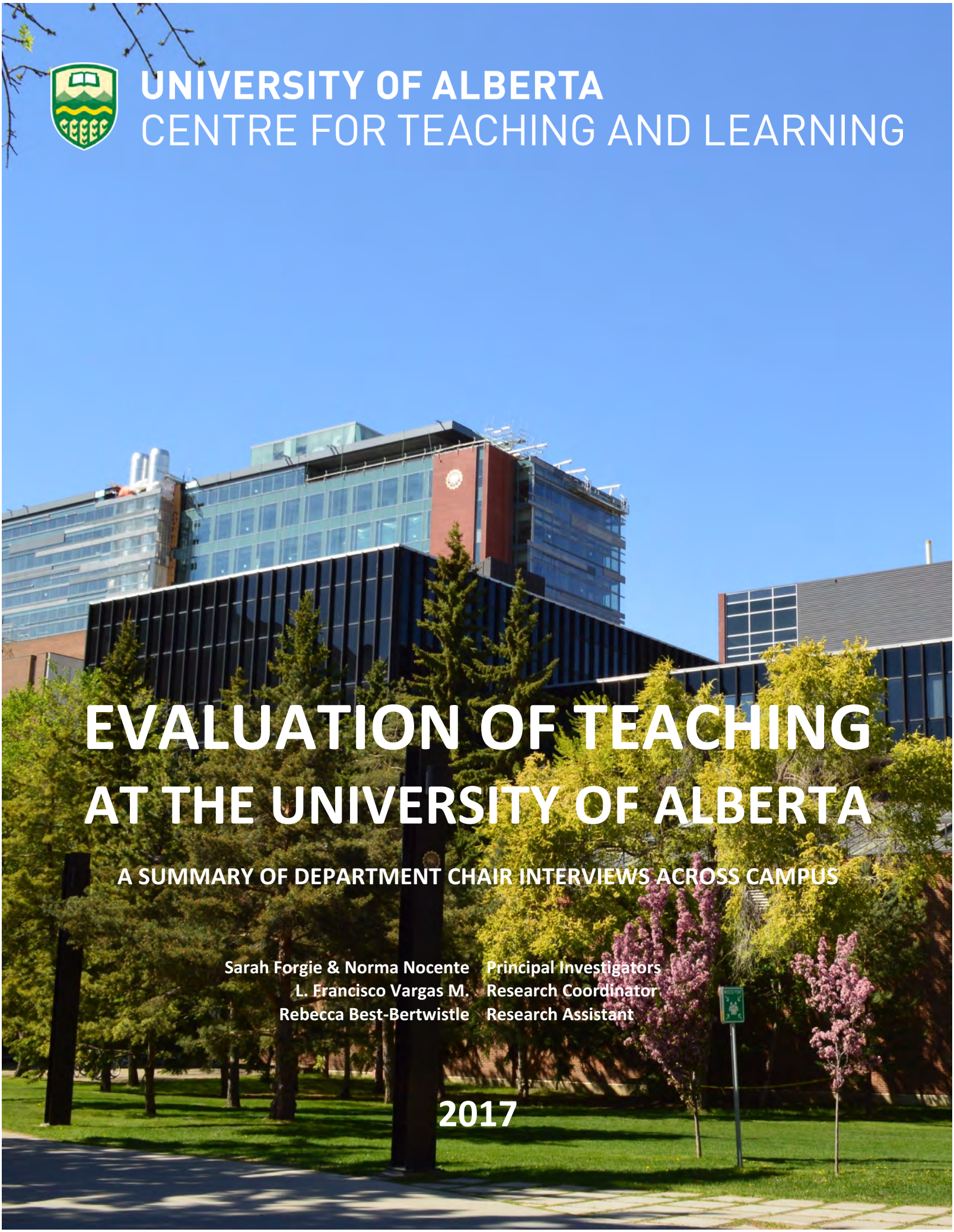
Marsh & Roche (1997): multidimensional aspects of teaching should be evaluated; suggest nine factors; “homemade” surveys are of questionable quality

Martin, Dennehy, & Morgan (2013): validity of SET is questioned; student focus groups suggested as an alternative

Ridley & Collins (2015): suggests a comprehensive performance evaluation instrument

Stupans, McGuren, & Babey (2016): present a tool for analyzing free-form comments on ratings forms

Zimmerman (2008): some tools may encourage students to focus on negative aspects of teaching; anonymous feedback means students are not accountable for their comments



UNIVERSITY OF ALBERTA
CENTRE FOR TEACHING AND LEARNING

EVALUATION OF TEACHING AT THE UNIVERSITY OF ALBERTA

A SUMMARY OF DEPARTMENT CHAIR INTERVIEWS ACROSS CAMPUS

Sarah Forgie & Norma Nocente — Principal Investigators
L. Francisco Vargas M. — Research Coordinator
Rebecca Best-Bertwistle — Research Assistant

2017

“I think these measures are useful, as long as they’re not used by themselves. They need to be supplemented by all kinds of other things” (Department Chair).



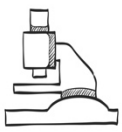


Table of Contents

1. Executive Summary

In May 2016, General Faculties Council tasked the Committee on Learning Environment to report on the "… research into the use of student rating mechanisms of instruction in university courses. This will be informed by a critical review of the University of Alberta's existing Universal Student Ratings of Instruction (USRIs) and their use for assessment and evaluation of teaching as well as a broad review of possible methods of multifaceted assessment and evaluation of teaching."

Methods

- Qualitative research. Department chairs (or their equivalents in non-departmental faculties) were asked to participate in short 30-45 minute (audio-recorded) semi-structured interviews with questions regarding their experiences evaluating teaching.
- Data was collected from January to March 2017, with a response rate of 59%.

Our committee sought to address the GFC motion by answering the following three questions:

1. What does the research have to say about student ratings of teaching?

    - A literature review on student rating systems previously presented in a 2009 University of Alberta report was updated (*Evaluation of Teaching at the U of A: Report of the Sub-Committee of the Committee on the Learning Environment*).

2. How are the USRIs and other tools used in the evaluation of teaching evaluation at the University of Alberta?

    - Participants from all faculties other than FOMD use USRI scores and comments (and only a portion of participants from FOMD) to evaluate teaching.
    - Statement 221 (overall the instructor was excellent), and statement 25 (overall the quality of the course content was excellent) are the most commonly used USRI items to evaluate teaching.
    - Most participants try to contextualize their interpretation of USRI results.

3. What are some approaches for multi-faceted evaluation of teaching?

    - In-class peer teaching observations were the most commonly used additional source of information, followed by annual instructor pedagogical self-reflections.
    - Most participants obtain these resources on a voluntary basis, only when professors agree to give them these supplementary resources.
    - Some participants have implemented yearly faculty audits, in which a manageable portion of their professorate's teaching is evaluated using additional information.
    - Even when participants obtain these resources, not all reported to bring them to FEC. When this information makes it to FEC, it is used to inform their narrative, and is only explicitly brought up when there is a concern with the numerical scores.
    - Despite more value being placed in teaching, most participants still described a strong bias towards research at their respective FECs.
        - Most participants voiced their need for additional supports to better evaluate teaching.
        - Most participants identified some issues when evaluating teaching exclusively with USRI, and some recommended possible alternatives to supplement these scores, but they still hope the institution will provide solutions for their concerns.

2. Introduction

The University of Alberta's Institutional Strategic Plan, For the Public Good, underscores its strong commitment to teaching and learning. The University community values the intellectual and engaging learning environment that is cultivated by our inspiring teachers. Accordingly, the evaluation of teaching is essential in upholding these values.

Teaching evaluations not only affect the careers of individuals at the University of Alberta, they also shape the quality of instruction being offered to students. Universal Student Ratings of Instruction (USRI) are often used to evaluate teaching quality for faculty annual review and tenure and promotion (summative evaluation). Also, USRIs can provide insight (formative evaluation) into specific areas of strength or improvement related to different aspects of teaching such as planning and organization, communication, assessment, etc. However, the dual purpose of USRIs is often contentious, particularly because of the perceived weight they carry with Faculty Evaluation Committees.

Consequently, in May 2016, General Faculties Council (GFC) tasked the Committee on Learning Environment (CLE) to report on the "… research into the use of student rating mechanisms of instruction in university courses. This will be informed by a critical review of the University of Alberta's existing Universal Student Ratings of Instruction (USRIs) and their use for assessment and evaluation of teaching as well as a broad review of possible methods of multifaceted assessment and evaluation of teaching. The ultimate objective will be to satisfy the Institutional Strategic Plan: For the Public Good strategy to: Provide robust supports, tools, and training to develop and assess teaching quality, using qualitative and quantitative criteria that are fair, equitable, and meaningful across disciplines."

Our committee sought to address the GFC motion by answering the following three questions:

1. What does the research have to say about student ratings of teaching?
2. How are the USRIs and other tools used in the evaluation of teaching evaluation at the University of Alberta?
3. What are some approaches for multi-faceted evaluation of teaching?

For the first question, we updated a literature review on student rating systems previously presented in a 2009 University of Alberta report (*Evaluation of Teaching at the U of A: Report of the Sub-Committee of the Committee on the Learning Environment).* To partially address the third question, we resurrected previous work completed at the University of Alberta on the multi-faceted evaluation of teaching. This information was presented to CLE in September 2016. This report primarily addresses the second and third question through information collected in interviews with department chairs across campus.

While University policy suggests that departments utilize a multi-faceted approach to evaluating teaching, we do not have a clear picture of the tools used other than the mandated Universal Student Rating System (USRI). These interviews helped to uncover how department chairs utilize USRIs to make personnel decisions and the helped to determine which other tools they used to evaluate the quality of teaching in their respective departments.

The purpose of this study is to describe the current state of teaching evaluation at the University of Alberta. More specifically it will help us understand the tools used to evaluate teaching at the University of Alberta.

3. Methods

Ethics approval for this qualitative study was sought from the Human Research Ethics Board at the University of Alberta, and obtained December 7, 2016 (Pro00069070). A qualitative approach with interviews was used to elicit the depth of response necessary for understanding the nuances and variety in possible answers.

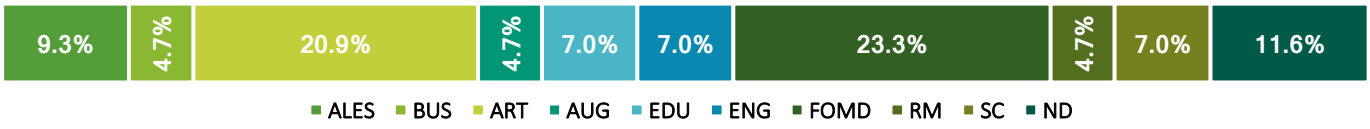
Department chairs (or their equivalents in non-departmental faculties) were emailed directly with information about the study, and with copy of the research letter of invitation. They were asked to participate in a short 30-45 minute (audio-recorded) semi-structured interview (*see Appendix 1*). The interview protocol was pre-approved by CLE, and it consisted of questions regarding the chairs' experiences evaluating teaching. Participants were also given two sample USRI case studies based on real teaching scores (*see Appendix 2*) and asked to interpret the scores. They were directed to reflect on both scores as if both instructors were teaching different sections of the same course within their department.

Data was collected from January to March 2017.

3.1. Participants

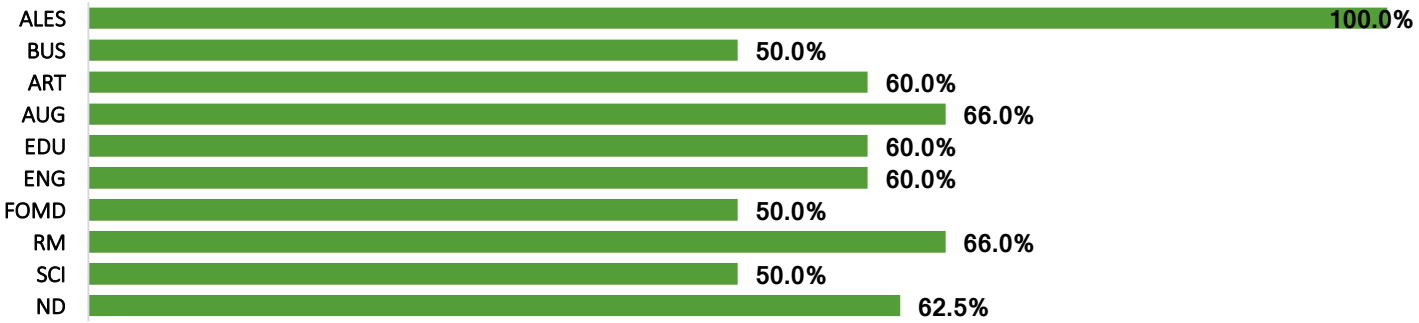
Participants were 43 department chairs (or their equivalents in non-departmental faculties) which is a 59% response rate. The distribution was 9.3% from Agricultural, Life and Environmental Sciences (ALES), 4.7% from Alberta School of Business (BUS), 20.9% from Arts (ART), 4.7% from Augustana Campus (AUG), 7% from Education (EDU), 7% from Engineering (ENG), 23.3% from Medicine and Dentistry (FOMD), 4.7% from Rehabilitation Medicine (RM), 7% from Science (SCI), and 11.6% from all non-departmental faculties (ND) (*see Figure 1*). Response rate reached a minimum of 50% within the different faculties (*see Figure 2*).

Figure 1. Distribution of Participants by Faculty



Participants reported having an average of 32.07 (*SD* = 22.42) faculty and FSO, 23.18 (*SD* = 27.03) sessional or contract instructors, and 3.06 (*SD* = 3.82) graduate students teaching in their departments. They mentioned working for an average of 4.34 (*SD* = 3.61) years as department chairs (or their equivalents in non-departmental faculties), and 9.3% of the total indicated having an interim appointment.

Figure 2. Response Rate by Faculty



3.2. Data Analysis

Confidentiality and anonymity were guaranteed by assigning pseudonyms to each audio file before it was sent for transcription. Transcripts were further anonymized by removing any information that identified the department under discussion (i.e., mention of disciplines, courses, specific individuals, and others). Participants from departmental faculties were grouped together and those from non-departmental faculties were combined to protect their identity. The complete list of participants, as well as assigned pseudonyms, is only available to the research coordinator. Interview transcripts were then coded with the qualitative data analysis software *NVivo 11*, using the main questions as general guidelines that informed the different codes/nodes. An external research assistant determined an inter-coder percentage agreement of .95 with 10% of the total number of interviews for the qualitative data, and of .98 with 100% of interviews for the quantitative representation of the data.

4. Results

This section offers both a quantitative and a qualitative summary of all participant responses, except section 4.1., section 4.2., and section 4.7., in which results only consider participants who reported using USRI. Information in these sections excludes participants from FOMD who indicated not using USRI, or whose application was not clear (*see Figure 3*).

4.1. Use of USRI to Evaluate Teaching

Figure 3. Participants from FOMD that Reported Using USRI Scores to Evaluate Teaching

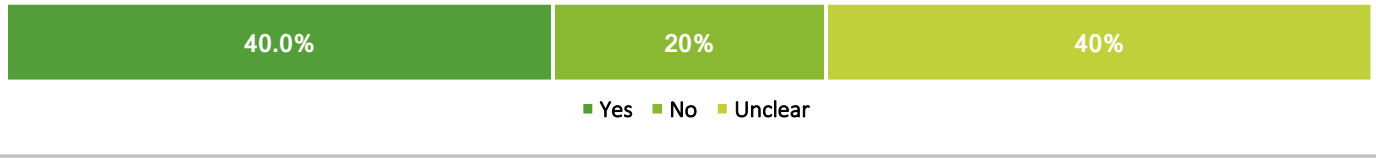


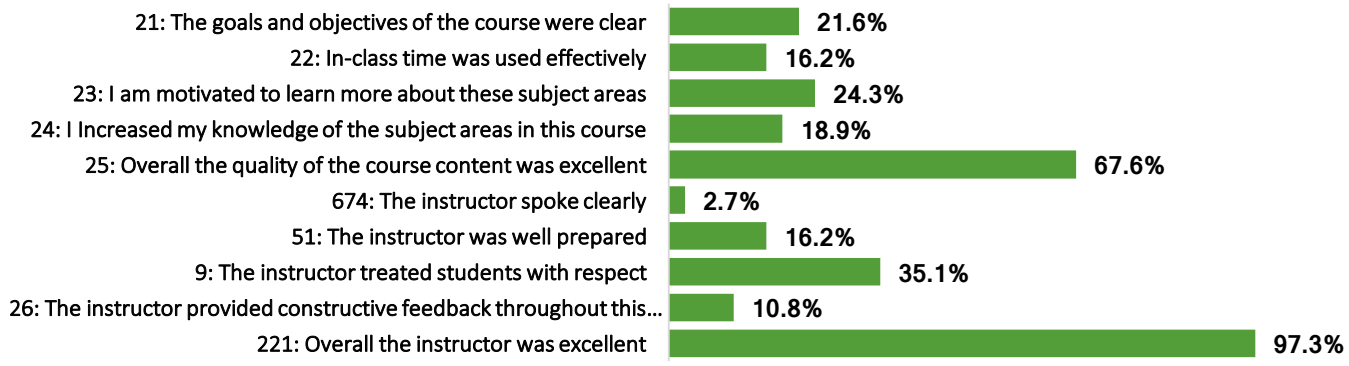
Figure 4. Participants from FOMD that Reported Using USRI Comments to Evaluate Teaching

Participants from all faculties other than FOMD reported using USRI scores and comments as part of their teaching evaluation process (100%). Department chairs from FOMD either mentioned using the USRI scores (40%), not using them (20%), or did not provide a definite answer (40%) (*see Figure 3*).

Additionally, department chairs from FOMD either indicated using USRI comments (30%), not taking them into consideration (30%), or their responses were unclear (40%) (*see Figure 4*). "*I have never seen it, but our largest undergraduate program has a different evaluation system, which is mainly based on narrative comments. So, your email, as I said, was the first time that I heard the term ever.*" They were often unsure if their department used USRI, or had never heard about USRI, or had never seen the scores (*see Appendix 2*).

FROM THIS POINT ON INFORMATION ONLY CONSIDERS PARTICIPANTS WHO REPORTED USING USRI

Figure 5. USRI Statements Most Commonly Used to Evaluate Teaching



When asked which USRI statements were most commonly used in their teaching evaluation process, statement 221 (overall this instructor was excellent) was identified by 97.3% of participants, statement 25 (overall the quality of the

course content was excellent) was selected by 67.6%, and statement 9 (the instructor treated students with respect) was identified by 35.1% (*see Figure 5*). In general, participants revealed that one or two items are used as an indicator of effective teaching. They seem to have benchmarks in mind as they review USRI scores:

> *We consider all of them, but of course we key in right away on 'the instructor was excellent.' You always look at that one first. And overall the course content was excellent is the second thing you look at. And then, if there's problems in either of those two scores you look in more detail at the other questions. There's something like 300 faculty members in the Faculty of Science for FEC, so we're only finding ways to efficiently go through these things.*

Figure 6. Participant Interpretation of Instructor A's USRI Scores

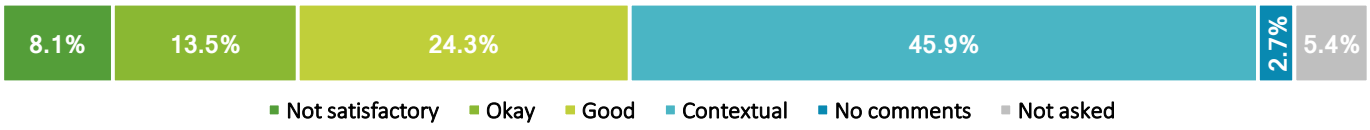
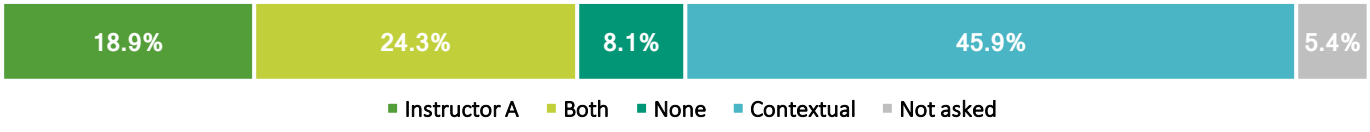


Figure 7. Participant Interpretation of Instructor B's USRI Scores

Participants also reflected on the USRI case studies (*see Appendix 2*). Instructor A had 6 USRI items on the 25th percentile or below, and 1 item below the Tukey fence. This instructor scored 4.0 on statement 221, 3.8 on statement 25, and 4.0 on statement 9. Instructor B had 7 USRI items between the 50th and 25th percentile, but no items were below the Tukey fence. This instructor scored 4.5 on statement 221, 4.2 on statement 25, and 4.8 on statement 9. After reflecting on these sample case studies, 8.1% of participants gave Instructor A 'unsatisfactory' reviews, 13.5% thought the scores were 'okay', and 24.3% considered the scores were 'good' (*see Figure 6*). Instructor B received more positive reviews, with 8.1% considering the scores were 'okay', 27% thinking they were 'good', and 10.8% deeming them as 'excellent' (*see Figure 7*). Moreover, believing the USRI data indicated their teaching was 'okay', 45.9% of participants mentioned that contextual factors should be considered in the evaluation of teaching (*see Figure 6 and 7*), and that to provide an informed interpretation of these USRI scores, they required more information than the one provided:

> *To be perfectly honest, in the abstract I don't know what I would say. Without knowing the circumstances, if one of those instructors is in her or his first year of teaching, and the other was an experienced professor, I think that interpretation is dramatically different than if they're both experienced professors or if they're both new professors. I can say, if we look at the overall averages they're both scoring in the lower percentile, and that sort of data, but to be perfectly honest that means very little to me because I think that understanding a person's position is crucial to being able to read any of these numbers.*

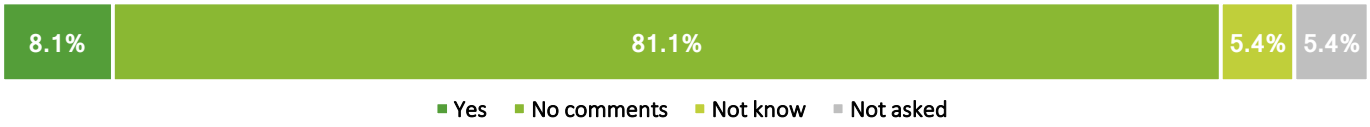
Figure 8. Participant Reported Case Studies Follow-Up



Additionally, 18.9% would only follow up with Instructor A to address issues related to their teaching scores, and/or to provide supplementary guidance to help them improve their results; 24.3% would follow up with both instructors to discuss their concerns; 8.1% would not follow up with either instructor, due to what they consider a lack of any teaching

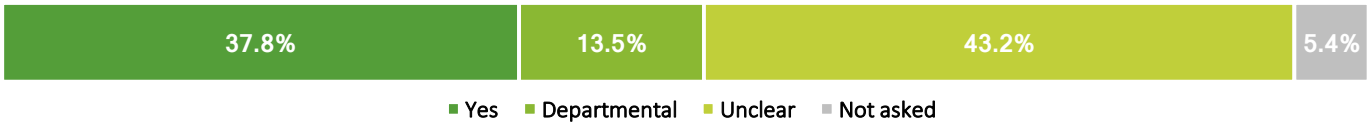
red flags; and 45.9% still mentioned that since USRI needs to be interpreted in a contextual way, they need to look into the circumstances of both instructors as part of their normal process (*see Figure 8*).

Figure 9. Participant Reference to Tukey Fence Data



Participants also had access to two pieces of reference data when given these case studies. The Tukey fence was not referenced by 81.1% of the participants, even though Instructor A had one score below the Tukey fence, and not all participants (5.4%) seemed familiar with its application (*see Figure 9*). The Test Scoring & Questionnaire Services (TSQS) Office mentioned that they generate diverse reports for different faculties and departments, and based on that, some participants might not be getting the complete set of data available. Participants were more familiar with quartiles data, however, as 37.8% of participants made explicit reference to them, 13.5% stated departmental expectations regarding USRI scores without making explicit reference to the quartiles, and 43.2% did not provide any definite comment (*see Figure 10*).

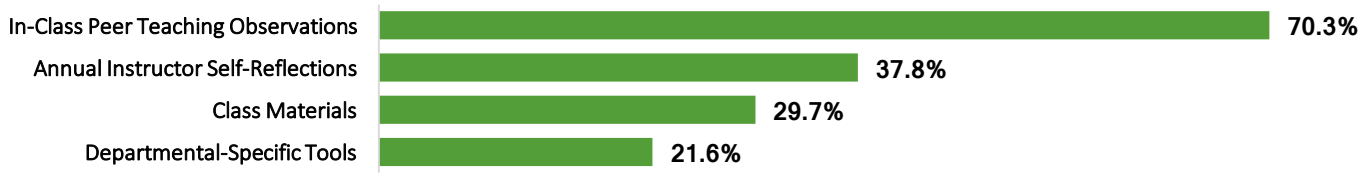
Figure 10. Participant Reference to Quartile Data



In general, participants from all faculties other than FOMD use USRI scores and comments (and only a portion of FOMD participants reported using this tool) to evaluate teaching. And even when one or two items are mainly used as an indicator of effective teaching, most participants try to contextualize their interpretations of USRI results.

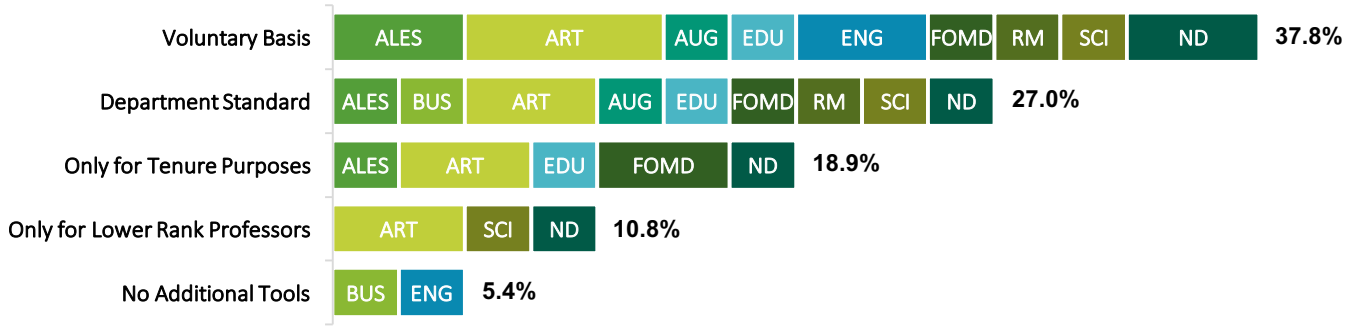
4.2. Use of Additional Tools & Information to Evaluate Teaching

Figure 11. Additional Tools & Information Most Commonly Used to Evaluate Teaching



When asked about the use of additional tools and information to evaluate teaching, in-class peer teaching observations were the most commonly implemented resource (70.3%), followed by annual instructor self-reflections about their pedagogical practices (37.8%), review of class materials (e.g., syllabi, assignments, and exams) (29.7%), and departmental specific tools that have been created to accommodate to the uniqueness of their departments (21.6%) (*see Figure 11*).

Figure 12. Distribution of Additional Tools & Information Use by Faculty

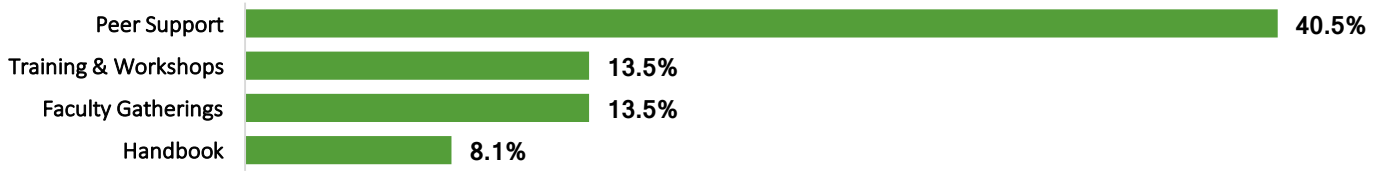


But the implementation of these tools varies between departments. Some participants (35.1%) only employ additional resources on a voluntary basis, encouraging professors to provide further information, but reportedly are not able to engage everyone in the department. Another group (27%) uses additional information as a standard, obtaining it through departmental specific tools. Some of them (8.1%) have already implemented yearly departmental audits that include additional tools and information. Furthermore, 18.9% only go beyond USRI when they need to evaluate teaching practices of professors going up for promotion/tenure; 10.8% only implement additional strategies to assess sessional instructors or new professors; and 8.1% acknowledged they did not use any additional tools or information (*see Figure 12*).

Among the participants who used additional tools and information in any way, 42.8% used one of the listed resources (*see Figure 11*), 42.8% used two, and 14.4% used three. Nevertheless, most participants share a common rationale for including other tools recognize the need to include other tools are very much alike, as one of them mentioned when reflecting on relying exclusively on USRI to evaluate teaching:

> *I don't think that's very useful by itself, it's incomplete. I'd feel uncomfortable judging somebody's fate just based on that. I'm not saying it's wrong but it's only one piece. It's one piece of understanding, and we take teaching seriously. It's not just a bunch of simple numbers pouring at us. We don't just look at you're above this number or below this number, and we're done. We're looking at you much more carefully than that, but it's a good start.*

Figure 13. Additional Tools & Information Used to Support Teaching



Participants, furthermore, mentioned tools and information they have utilized in their departments to *support* teaching. For instance, 40.5% have organized peer support initiatives (e.g., mentoring, teaching triads, and support groups where instructors find a safe space to talk about their teaching practices). Another 13.5% have referred struggling faculty to departmental specific training and/or workshops, or to other units on campus that offer pedagogical guidance; 13.5% have instituted faculty gatherings to open casual conversations about teaching practices and problems. Additionally, 8.1% have produced departmental teaching handbooks (*see Figure 13*).

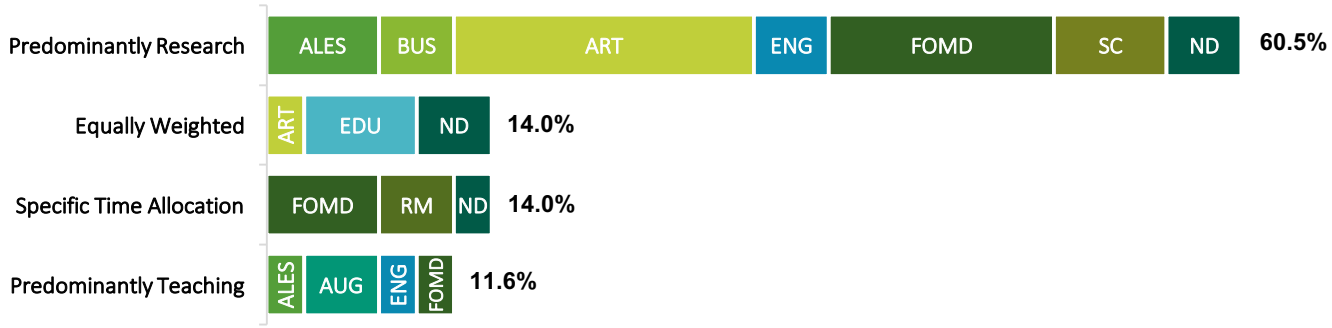
Figure 14. Percentage of Participants that Bring Additional Tools & Information to FEC

When it comes to bringing this additional tools and information to FEC, 45.9% indicated that these sources play a role in their annual teaching evaluation, by informing a narrative and/or the reasoning with other FEC members if their recommendation gets challenged; 21.6% acknowledged not bringing these resources to FEC, and 32.4% did not comment or their responses were unclear (*see Figure 14*). Thus, even when participants indicated using one or two additional tools to evaluate teaching, most acknowledged using them on a voluntary basis, receiving this information only when faculty agrees to provide these supplementary resources.

4.3. Perceived FEC Weighting of Teaching, Research & Service

FROM THIS POINT ON INFORMATION CONSIDERS ALL PARTICIPANTS

Figure 15. Distribution of Perceived FEC Weighting



Most participants recognized that there is a strong bias towards research (60.5%), despite their FEC's best efforts to weight them equally (14%) (*see Figure 19*):

> *I would say that there's still a bias towards research. Although my experience was that teaching was taken seriously, and we looked at those things a lot, and they were raised in terms of the kinds of things people were doing, the amount of teaching they were doing, their scores, and all that stuff was taken into consideration, I would still say that the publications and other research activities and outcomes were probably weighed more seriously. So, I'd say it'd be more like 50%, 30%, 20% rather than 40%, 40%, 20%.*

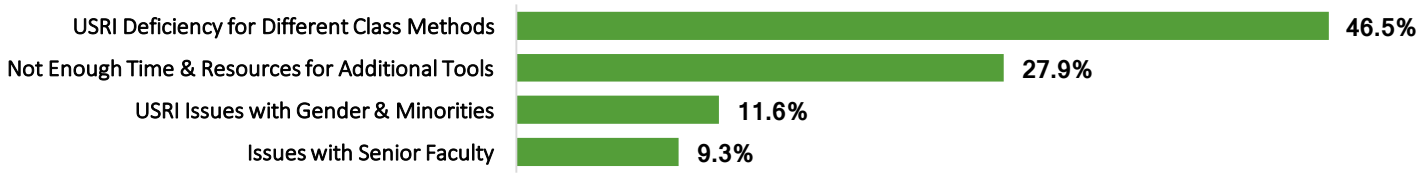
An additional 14% mentioned that FEC weights the importance of teaching, research and service based on the specific time allocation of the individual (mostly in health-related disciplines where their contracts have different time allocations), and 11.6% thought that their FEC weights teaching more heavily than research (*see Figure 15*).

4.4. Need for Additional Supports to Better Evaluate Teaching

Figure 16. Perceived Need for Additional Supports to Better Evaluate Teaching

Most participants also voiced their urgent need for additional supports to better evaluate teaching. One participant, for example, remarked that "*I was looking to you to find this out, to find out if the result of this survey would give me some ideas of what this is*"; and another commented that in their department "*We're hoping the university will solve this issue.*" Indeed, 83.7% of participants mentioned needing some support, whereas 9.3% indicated not needing additional resources (*see Figure 16*).

Figure 17. Issues Encountered when Evaluating Teaching

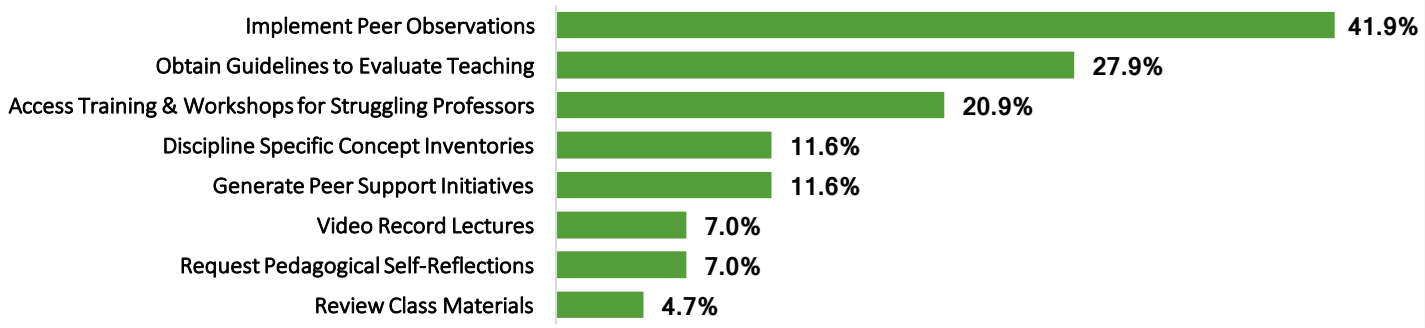


Some participants explicitly recognized their concerns about depending exclusively on USRI, and the inability of USRIs to effectively evaluate diverse approaches to teaching (46.5%), other mentioned not having enough time and resources to adopt supplementary tools in the teaching evaluation process (27.9%). Participants also expressed concerns about lower USRI scores for women and visible minorities (11.6%), as well as the difficulties of compelling senior faculty (usually with full professor rank) to improve their teaching practices (9.3%) (*see Figure 17*):

> *That question set doesn't serve the diversity and the kind of pedagogy we have now, and really needs fixing. I think there needs to be a conversation about what this is going to look like over time. I also think the University has to take very seriously the concerns that equity seeking groups have about what happens in teaching evaluations. What happens to women? What happens to visible minority? What happens to people that are perceived to have strong accents? And I think there's a huge responsibility on chairs and people on FEC to really be educated in how much you can extrapolate from USRI.*

TSQS conducted descriptive analyses that generated gender-specific USRI scores using data from the academic years 2011/2012 to 2015/2016. Results show there is no overt difference between scores for males (N = 18576, Mdn = 4.53) and females (N = 13679, Mdn = 4.57) for statement 211. Additionally, TSQS measures the reliability of the USRI by comparing medians to the previous academic years. Our research team was not able to find information on the validity of the USRI.

Figure 18. Most Common Ideal Types of Supports to Better Evaluate Teaching



Among the most commonly listed types of supports to better evaluate teaching, participants mentioned that ideally, they would implement peer in-class observations not only for promotion purposes, but across their department (41.9%), obtain university guidelines to understand how to accurately and effectively evaluate teaching (27.9%) (*see Figure 18*):

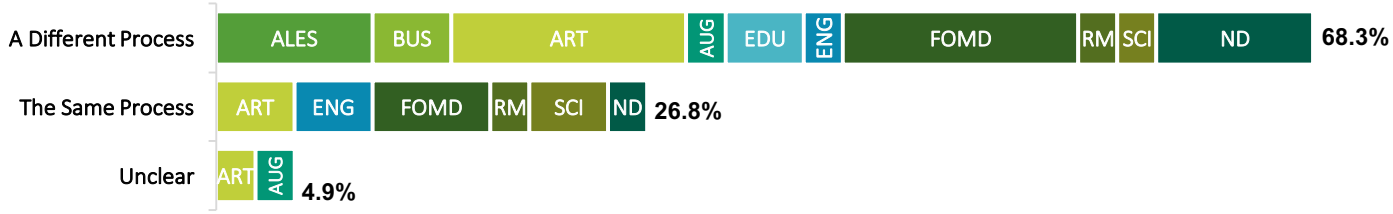
> *My learning curve coming in to the chair role has been huge. We used to have a chair's school kind of thing. Now there's the gold and green leadership college or whatever it's called, and it's a very different thing. So, you transition into chair now and you're on your own. You've got to go figure it out, ask people for coffee, and learn up, but there's no orientation to being a chair.*

Some also indicated that it would be useful to gain access to teaching training and workshops that they could refer struggling professors to (when not available in their departments) (20.9%), have discipline specific concept inventories to better determine the knowledge increase in students (11.6%), implement peer support initiatives to improve teaching practices (11.6%), video record lectures for later analysis of the quality of teaching (7%), request pedagogical self-reflections in which professors give a thoughtful summary of their teaching (7%), and review class materials to have a better panorama of the instructor (4.7%) (*see Figure 18*). Having more resources to better evaluate teaching is important, as one of them mentioned:

> *I think we need support to develop our own teaching skills more comfortably so we can be excellent teachers, but also it would be important to make sure our instruments are valid and that we can actually use them on a journey of self-improvement, and departmental culture and improvement. And to do that having some facilitation from people who know the art and who can work with us would be better than just having a list of stuff on a website where you do click, click, and access what you want. That's not enough.*

4.5. Difference Between Teaching Evaluation for Annual Review & Promotion

Figure 19. Perceived Differences between Evaluation of Teaching for Annual Review & Promotion

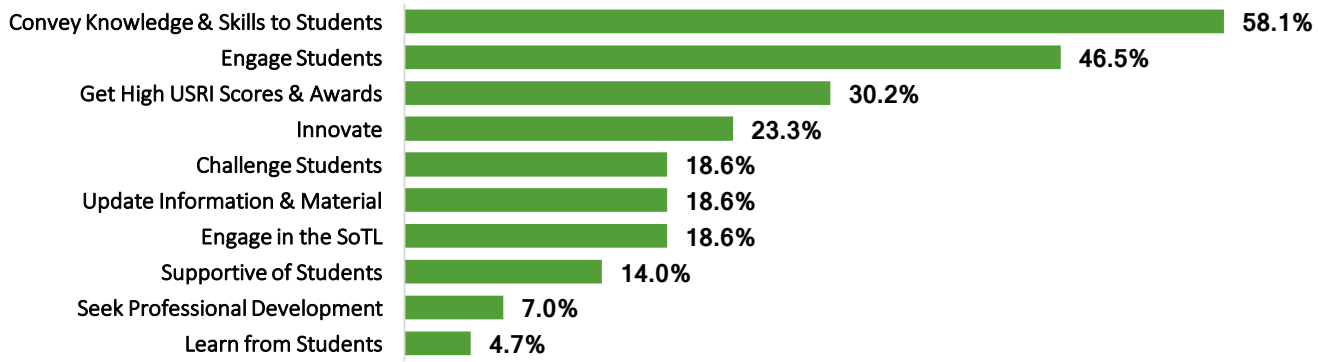


Even though evaluation of teaching for annual review and for promotion was a different process for 68.3%, and the same process for 26.8% of participants (*see Figure 19*), both ends of the spectrum seem to agree that more components were taking into consideration when they were dealing with promotion:

> *The annual review looks only at that year, and if there's real concerns then you'll look for trends, whereas when it comes to promotion, it looks to a career, what has this individual been doing with teaching, and not just this year but intentionally over the entire career. When it comes to application promotion, there is a larger view taken of teaching*.

4.6. Characteristics of Effective & Excellent Teachers

Figure 20. Characteristics of Effective & Excellent Teachers



Even though most participants struggled with the breadth of this question, for them an effective and/or excellent teacher appropriately conveys the knowledge and the skills that students need to obtain (58.1%), engages students despite the difficulty of the course material (46.5%), gets high USRI scores and teaching awards (30.2%), innovates in their teaching practices (23.3%), knows how to challenge students without burning them out (18.6%), regularly updates the information and the material of the course (18.6%), and engages in scholarship of teaching and learning related activities (18.6%). Other participants indicated that being supportive of students was also important (14%), seeking professional development opportunities to improve their pedagogical practices (7%), and learning from students as much as students learn from them (4.7%) (*see Figure 21*):

> *I try to avoid definitions if that involves any kind of explicit criteria. What I look for, what I think is most important in teaching is that all good teaching is transformative. And it's mostly transformative for the student, although truth be known good teaching is transformative for both student and teacher*.

4.7. Experiences Transitioning to e-USRI Compared to Paper-Based USRI

FROM THIS POINT ON INFORMATION ONLY CONSIDERS PARTICIPANTS WHO REPORTED USING USRI

Figure 21. Reported Response Rate Experiences with e-USRI compared to Paper-Based USRI

Most participants believed that response rates have decreased since the implementation of the e-USRI: 48.6% had some data to back up this claim, such as their personal USRI response rates, or the actual number of students that now complete the evaluations compared to previous years; and 18.9% believed that the response rates had declined, but had no data to support this claim. Alternatively, 21.6% of participants believed there was a similar response rate with both methods of delivery, 8.1% thought that it increased with the switch to electronic, but did not offer data to support this claim (*see Figure 21*). Moreover, some participants (8.1%) believed that a major issue with USRI response rates is that students are asked to complete a large amount of assessments:

> *I think they get completely annoyed because they're being bombarded with e-mails in their last week of classes reminding them to do USRIs, and professors reminding them to do USRIs to the point where I think they just go: I'm really annoyed. I'm not going to do them at all. I don't know what kind of a system they use to send them out, but it's almost like they send out one for every class, for every student, so they're just harassing them to death and they get mad about it*.

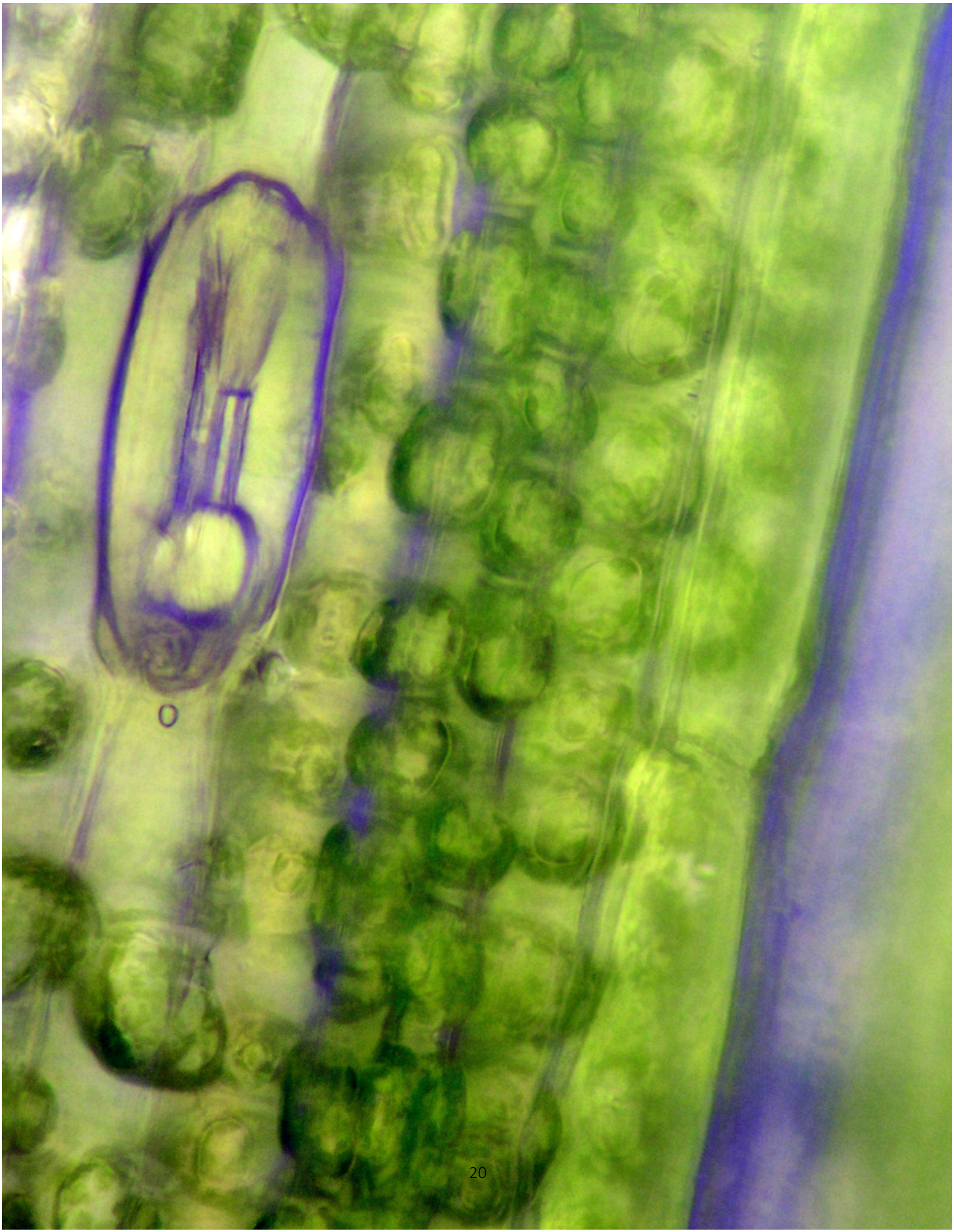
5. Conclusions

How are the USRIs and other tools used in the evaluation of teaching evaluation at the University of Alberta?

- Participants from all faculties other than FOMD use USRI scores and comments (and only a portion of participants from FOMD) to evaluate teaching.
- Statement 221 (overall the instructor was excellent), and statement 25 (overall the quality of the course content was excellent) are the most commonly used USRI items to evaluate teaching.
- Most participants try to contextualize their interpretation of USRI results.

What are some approaches for multi-faceted evaluation of teaching?

- In-class peer teaching observations were the most commonly used additional source of information, followed by annual instructor pedagogical self-reflections.
- Most participants obtain these resources on a voluntary basis, only when professors agree to give them these supplementary resources.
- Some participants have implemented yearly faculty audits, in which a manageable portion of their professorate's teaching is evaluated using additional information.
- Even when participants obtain these resources, not all reported to bring them to FEC. When this information makes it to FEC, it is used to inform their narrative, and is only explicitly brought up when there is a challenge.
- Participants recognized that there is still a strong bias towards research at their respective FEC.
    - Most participants voiced their need for additional supports to better evaluate teaching.
    - They have identified some issues when evaluating teaching exclusively with USRI, and possible alternatives to supplement these scores, but still they hope the institution provides a solution for their concerns.


O

6. Appendix 1: Semi-Structured Interview Questions

Study Title: Evaluation of Teaching at the University of Alberta

1. Demographics
    a. Identify department/faculty
    b. Number of faculty/ FSOs who teach
    c. Number of sessionals who teach
    d. Number of graduate students who teach

2. How do you evaluate teaching?

    a. Do you (or your FEC) use USRIs to evaluate the teaching of your faculty members?

    b. If yes, which of the following standard USRI statements are considered in your faculty's teaching evaluation process?

        i. the goals and objectives of the course were clear
        ii. in-class time was used effectively
        iii. I am motivated to learn more about these subject areas
        iv. I increased my knowledge of the subject areas in this course
        v. Overall the quality of the course content was excellent
        vi. the instructor spoke clearly
        vii. the instructor was well prepared
        viii. the instructor treated students with respect
        ix. the instructor provided constructive feedback throughout this course
        x. overall this instructor was excellent

3. How do you compare your experience with e-USRIs and in-class paper-based USRIs?

4. What, if any, additional tools do you regularly use, other than USRI to evaluate teaching? If you don't, why not?

5. Do you use additional sources of information to evaluate teaching? If so, what information do you use and how are these sources of information weighted in teaching evaluations? Why?

6. Do you believe most of the FEC members weight teaching, research and service equally? If not, describe the average weighting, in your opinion.

7. How is evaluation of teaching different (or not) for annual review, or for promotion?

8. How do you define effective and/or excellent teaching? Do you have set standards, or do you make a relative comparison?

9. What additional supports would be useful to you to better evaluate teaching?

7. Appendix 2: Sample USRI Results for Department Chairs

Study Title: Evaluation of Teaching at the University of Alberta

Please look at the USRI information provided for two different instructors teaching the same course. How would you describe the instructors' teaching to FEC? OR In terms of evaluating teaching, what is your interpretation of this data for each instructor?

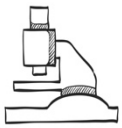
Instructor A

		Reference Data			
Question	Median	Tukey Fence	25%	50%	75%
The goals and objectives of the course were clear	3.4	2.7	3.9	4.3	4.7
In-class time was used effectively.	3.6	2.5	3.8	4.3	4.7
I am motivated to learn more about these subject areas.	3.5	2.9	4.1	4.5	4.8
I increased my knowledge of the subject areas in this course.	4.4	3.0	4.1	4.6	4.8
Overall, the quality of the course content was excellent.	3.8	2.4	3.8	4.3	4.8
The instructor spoke clearly.	4.5	3.8	4.5	4.8	4.9
The instructor was well prepared.	4.6	3.4	4.3	4.8	4.9
The instructor treated the students with respect.	4.0	4.2	4.7	4.9	5.0
The instructor provided constructive feedback throughout this course.	4.5	2.8	4.0	4.5	4.8
Overall, this instructor was excellent.	4.0	3.2	4.2	4.7	4.9

Instructor B

		Reference Data			
Question	Median	Tukey Fence	25%	50%	75%
The goals and objectives of the course were clear	4.0	2.7	3.9	4.3	4.7
In-class time was used effectively.	4.2	2.5	3.8	4.3	4.7
I am motivated to learn more about these subject areas.	3.7	2.9	4.1	4.5	4.8
I increased my knowledge of the subject areas in this course.	4.1	3.0	4.1	4.6	4.8
Overall, the quality of the course content was excellent.	4.2	2.4	3.8	4.3	4.8
The instructor spoke clearly.	4.7	3.8	4.5	4.8	4.9
The instructor was well prepared.	4.4	3.4	4.3	4.8	4.9
The instructor treated the students with respect.	4.8	4.2	4.7	4.9	5.0
The instructor provided constructive feedback throughout this course.	4.0	2.8	4.0	4.5	4.8
Overall, this instructor was excellent.	4.5	3.2	4.2	4.7	4.9





UNIVERSITY OF ALBERTA
CENTRE FOR TEACHING AND LEARNING

Appendix C: Interview Questions

Study Title: **Evaluation of Teaching at the University of Alberta**

1. Demographics
    a. Identify department/faculty
    b. Number of faculty/ FSOs who teach
    c. Number of sessionals who teach
    d. Number of graduate students who teach

2. How do you evaluate teaching?
    a. Do you (or your FEC) use USRIs to evaluate the teaching of your faculty members?
    b. If yes, which of the following standard USRI statements are considered in your faculty's teaching evaluation process?

        i. the goals and objectives of the course were clear
        ii. in-class time was used effectively
        iii. I am motivated to learn more about these subject areas
        iv. I increased my knowledge of the subject areas in this course
        v. Overall the quality of the course content was excellent
        vi. the instructor spoke clearly
        vii. the instructor was well prepared
        viii. the instructor treated students with respect
        ix. the instructor provided constructive feedback throughout this course
        x. overall this instructor was excellent

3. How do you compare your experience with e-USRIs and in-class paper-based USRIs?
4. What, if any, additional tools do you regularly use, other than USRI to evaluate teaching? If you don't, why not?
5. Do you use additional sources of information to evaluate teaching? If so, what information do you use and how are these sources of information weighted in teaching evaluations? Why?
6. Do you believe most of the FEC members weight teaching, research and service equally? If not, describe the average weighting, in your opinion.
7. How is evaluation of teaching different (or not) for annual review, or for promotion?
8. How do you define effective and/or excellent teaching? Do you have set standards, or do you make a relative comparison?
9. What additional supports would be useful to you to better evaluate teaching?

Appendix D: Sample USRI Case Studies

Study Title: **Evaluation of Teaching at the University of Alberta**

Please look at the USRI information provided for two different instructors teaching the same course. How would you describe the instructors' teaching to FEC? OR In terms of evaluating teaching, what is your interpretation of this data for each instructor?

Instructor A

		Reference Data			
Question	**Median**	**Tukey Fence**	**25%**	**50%**	**75%**
The goals and objectives of the course were clear	3.4	2.7	3.9	4.3	4.7
In-class time was used effectively.	3.6	2.5	3.8	4.3	4.7
I am motivated to learn more about these subject areas.	3.5	2.9	4.1	4.5	4.8
I increased my knowledge of the subject areas in this course.	4.4	3.0	4.1	4.6	4.8
Overall, the quality of the course content was excellent.	3.8	2.4	3.8	4.3	4.8
The instructor spoke clearly.	4.5	3.8	4.5	4.8	4.9
The instructor was well prepared.	4.6	3.4	4.3	4.8	4.9
The instructor treated the students with respect.	4.0	4.2	4.7	4.9	5.0
The instructor provided constructive feedback throughout this course.	4.5	2.8	4.0	4.5	4.8
Overall, this instructor was excellent.	4.0	3.2	4.2	4.7	4.9

Instructor B

		Reference Data			
Question	**Median**	**Tukey Fence**	**25%**	**50%**	**75%**
The goals and objectives of the course were clear	4.0	2.7	3.9	4.3	4.7
In-class time was used effectively.	4.2	2.5	3.8	4.3	4.7
I am motivated to learn more about these subject areas.	3.7	2.9	4.1	4.5	4.8
I increased my knowledge of the subject areas in this course.	4.1	3.0	4.1	4.6	4.8
Overall, the quality of the course content was excellent.	4.2	2.4	3.8	4.3	4.8
The instructor spoke clearly.	4.7	3.8	4.5	4.8	4.9
The instructor was well prepared.	4.4	3.4	4.3	4.8	4.9
The instructor treated the students with respect.	4.8	4.2	4.7	4.9	5.0
The instructor provided constructive feedback throughout this course.	4.0	2.8	4.0	4.5	4.8
Overall, this instructor was excellent.	4.5	3.2	4.2	4.7	4.9

Appendix G: References for Reviewed Literature

These are the references used to review the literature only. Other references consulted for preparation of the report (such as University of Alberta reports and documents) are included at the end of the report.

Al-Eidan, F., Baig, L. A., Magzoub, M., & Omair, A. (2016). Reliability and validity of the faculty evaluation instrument used at King Saud bin Abdulaziz University for Health Sciences: Results from the haematology course. *The Journal of the Pakistan Medical Association, 66*(4), 453-457. http://www.jpma.org.pk/full_article_text.php?article_id=7711

Backer, E. (2012). Burnt at the student evaluation stake – the penalty for failing students. *E-Journal of Business Education & Scholarship of Teaching, 6*(1), 1-13. Retrieved from http://www.ejbest.org/upload/eJBEST_Backer_2012_1.pdf

Bedggood, R. E., & Donovan, J. D. (2012). University performance evaluations: What are we really measuring? *Studies in Higher Education, 37*(7), 825-842. http://dx.doi.org/10.1080/03075079.2010.549221

Berk, R. A. (2013). Top five flashpoints in the assessment of teaching effectiveness. *Medical Teacher, 35*(1), 15-26. http://dx.doi.org/10.3109/0142159X.2012.732247

Blackhart, G. C., Peruche, B. M., DeWall, C. N., & Joiner, T. E., Jr. (2006). Faculty forum: Factors influencing teaching evaluations in higher education. *Teaching of Psychology, 33*(1), 37-39. http://dx.doi.org/10.1207/s15328023top3301_9

Blair, E., & Valdez Noel, K. (2014). Improving higher education practice through student evaluation systems: is the student voice being heard? *Assessment & Evaluation in Higher Education, 39*(7), 879-894. http://dx.doi.org/10.1080/02602938.2013.875984

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

Boysen, G.A. (2015). Uses and misuses of student evaluations of teaching: The interpretation of differences in teaching evaluation means irrespective of statistical information. *Teaching of Psychology, 42*(2), 109-118. http://dx.doi.org/10.1177/0098628315569922

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2014). The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education, 39*(6), 641-656. http://dx.doi.org/10.1080.02602938.2013.860950

Brown, G. D. A., Wood, A. M., Ogden, R. S., & Maltby, J. (2014). Do student evaluations of university reflect inaccurate beliefs or actual experience? A relative rank model. *Journal of Behavioral Decision Making, 28*, 14-26. http://dx.doi.org/10.1002/bdm.1827

Campbell, J. P., & Bozeman, W. C. (2008). The value of student ratings: Perceptions of students, teachers, and administrators. *Community College Journal of Research and Practice, 32*, 13-24. http://dx.doi.org/10.1080/10668920600864137

Centra, J.A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*(5), 495-518. http://www.jstor.org.login.ezproxy.library.ualberta.ca/stable/40197319

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching?

The Journal of Higher Education, 71(1), 17-44. http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsjsr&AN=edsjsr.10.2307.2649280&site=eds-live&scope=site

Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: an assessment of student perception and motivation. *Assessment & Evaluation in Higher Education, 28*(1), 71-88. http://dx.doi.org/10.1080/0260293032000033071

Cheng, D. A. (2015). Effects of professorial tenure on undergraduate ratings of teaching performance. *Education Economics, 23*(3), 338-357. http://dx.doi.org/10.1080/09645292.2013.826632

Cho, D., Baek, W., & Cho, J. (2015). Why do good performing students highly rate their instructors? Evidence from a natural experiment. *Economics of Education Review, 49*, 172-179. http://dx.doi.org/10.1016/j.econedurev.2015.10.001

Cho, J., & Otani, K. (2014). Differences in student evaluations of limited-term lecturers and full-time faculty. *Journal on Excellence in College Teaching, 25*(2), 5-24. http://opus.ipfw.edu/profstudies_facpubs/64

Chonko, L. B., Tanner, J. F., & Davis, R. (2002). What are they thinking? Students' expectations and self-assessments. *Journal of Education for Business, 77*(5), 271-281. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=7214031&site=eds-live&scope=site

Clayson, D. E. (2013). Initial impressions and the student evaluation of teaching. *Journal of Education for Business, 88*(1), 26-53. http://dx.doi.org/10.1080/08832323.2011.633580

Cohen, E. H. (2005). Student evaluations of course and teacher: factor analysis and SSA approaches. *Assessment & Evaluation in Higher Education, 30*(2), 123-136. http://dx.doi.org/10.1080/0260293042000264235

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*(3), 281-309.

Cox, C.D., Peeters, M. J., Stanford, B. L., & Seifert, C. F. (2013). Pilot of peer assessment within experiential teaching and learning. *Currents in Pharmacy Teaching and Learning, 5*(4), 311-320. http://dx.doi.org/10.1016/j.cptl.2013.02.003

Curwood, J.S., Tomitsch, M., Thomson, K., & Hendry. G.D. (2015). Professional learning in higher education: Understanding how academics interpret student feedback and access resources to improve their teaching. *Australasian Journal of Educational Technology, 31*(5). http://dx.doi.org/10.14742/ajet.2516

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*(11), 1198-1208. http://dx.doi.org/10.1037/0003-066X.52.11.1198

Dolmans, D. M., Janssen-Noordman, A., & Wolfhagen, H. P. (2006). Can students differentiate between PBL tutors with different tutoring deficiencies? Medical Teacher, 28(6), 156-161. doi: 10.1080/01421590600776545

Dodeen, H. (2013). Validity, reliability, and potential bias of short forms of students' evaluation of teaching: The case of UAE University. *Educational Assessment, 18*(4), 235-250. http://dx.doi.org/10.1080/10627197.2013.846670

Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: the

relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education, 29*(1), 91-108. http://dx.doi.org/10.1080/0260293032000158180

Fraile, R., & Bosch-Morell, F. (2015). Considering teaching history and calculating confidence intervals in student evaluations of teaching quality: An approach based on Bayesian inference. *Higher Education, 70*(1), 55-72. http://dx.doi.org/10.1007/s10734-014-9823-0

Gehrt, K., Louie, T. A., & Osland, A. (2015). Student and professor similarity: Exploring the effects of gender and relative age. *Journal of Education for Business, 90*, 1-9. http://dx.doi.org/10.1080/08832323.2014.968514

Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: the perspective of currently enrolled students. *Studies in Higher Education, 32*(5), 603-615. http://dx.doi.org/10.1080/03075070701573773

Grammatikopoulos, V., Linardakis, M., Gregoriadis, A., & Oikonomidis, V. (2015). Assessing the students' evaluations of educational quality (SEEQ) questionnaire in Greek higher education. *Higher Education, 70*(3), 395-408. http://dx.doi.org/10.1007/s10734-014-9837-7

Grayson, J. P. (2015). Repeated low teaching evaluations: A form of habitual behavior? *Canadian Journal of Higher Education, 45*(4), 298-321. http://journals.sfu.ca/cjhe/index.php/cjhe/article/view/184404

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182-1186. http://dx.doi.org/10.1037/0003-066X.52.11.1182

Greenwald, A. G., Gillmore, G. M. (1997). Grade leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209-1217. http://dx.doi.org/10.1037/0003-066X.52.11.1209

Greimel-Fuhrmann, B. (2014). Student's perception of teaching behaviour and its effect on evaluation. *International Journal for Cross-Disciplinary Subjects in Education, 5*(1), 1557-1563. http://dx.doi.org/10.20533/ijcdse.2042.6364.2014.0218

Gump, S.E. (2007). Student evaluations of teaching effectiveness and the leniency hypothesis: A literature review. *Education Research Quarterly, 30*(3), 55-68. Retrieved from http://eric.ed.gov.login.ezproxy.library.ualberta.ca/?id=EJ787711

Huebner, L., & Magel, R. C. (2015). A gendered study of student ratings of instruction. *Open Journal of Statistics, 5,* 552-567. http://dx.doi.org/10.4236/ojs.2015.56058

Hughes II, K. E., & Pate, G. R. (2013). Moving beyond student ratings: A balanced scorecard approach for evaluating teaching performance. Issues in *Accounting Education, 28*(1), 49-75. http://dx.doi.org/10.2308/iace-50302

Iqbal, I. (2013). Academics' resistance to summative peer review of teaching: questionable rewards and the importance of student evaluations. *Teaching in Higher Education, 18*(5), 557-569. http://dx.doi.org/10.1080/13562517.2013.764863

Jackson, M. J., & Jackson, W. T. (2015). The misuse of student evaluations of teaching: Implications, suggestions and alternatives. *Academy of Educational Leadership Journal, 19*(3), 165-173. http://www.alliedacademies.org/academy-of-educational-leadership-journal/

Jones, J., Gaffney-Rhys, R., & Jones, E. (2014). Handle with care! An exploration of the

potential risks associated with the publication and summative usage of student evaluation of teaching (SET) results. *Journal of Further and Higher Education, 38*(1), 37-56. http://dx.doi.org/10.1080/0309877X.2012.699514

Keeley, J. W., English, T., Irons, J., & Henslee, A. M. (2013). Investigating halo and ceiling effects in student evaluations of instruction. *Educational and Psychological Measurement, 73*(3), 440-457. http://dx.doi.org/10.1177/0013164412475300

Khong, T. L. (2014). The validity and reliability of the student evaluation of teaching: A case in a private higher educational institution in Malaysia. *International Journal for Innovation Education and Research, 2*(9), 57-63. http://www.ijier.net/index.php/ijier/article/view/317

Kim, L. E., MacCann, C. (2016). What is students' ideal university instructor personality? An investigation of absolute and relative personality preferences. *Personality and Individual Differences, 102*, 190-203. http://dx.doi.org/10.1016/j.paid.2016.06.068

Kuwaiti, A. A., AlQuraan, M., & Subbarayalu, A. V. (2016). Understanding the effect of response rate and class size interaction on students evaluation of teaching in a higher education. *Educational Assessment & Evaluation, 3*, https://doi.org/10.1080/2331186X.2016.1204082

Lama, T., Arias, P., Mendoza, K. & Manahan, J. (2015). Student evaluation of teaching surveys: do students provide accurate and reliable information? *e-Journal of Social & Behavioural Research in Business, 6*(1), 30-39. http://www.ejsbrb.org/a.php?/content/issue/10

Laube, H., Massoni, K., Sprague, J., & Ferber, A. L. (2007). The impact of gender on the evaluation of teaching: What we know and what we can do. *NWSA Journal, 19*(3), 87-104. Retrieved from http://www.jstor.org/stable/40071230

Lyde, A.R., Grieshaber, D.C., Byrns, G. (2016). Faculty teaching performance: Perceptions of a multi-source method for evaluation (MME). *Journal of the Scholarship of Teaching and Learning, 16*(3), 82-94. http://dx.doi.org/10.14434/josotl.v16i3.18145

Macfadyen, L. P., Dawson, S., Prest, S., & Gasevic, D. (2016). Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations. *Assessment & Evaluation in Higher Education, 41*(6), 821-839. http://dx.doi.org/10.1080/02602938.2015.1044421

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. http://dx.doi.org/10.1007/s10755-014-9313-4

Makondo, L., & Ndebele, C. (2014). University lecturers' views on student-lecturer evaluations. *Anthropologist, 17*(2), 377-386. http://www.krepublishers.com/02-Journals/T-Anth/Anth-17-0-000-14-Web/Anth-17-0-000-14-Contents/Anth-17-0-000-14-Contents.htm

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187-1197. http://dx.doi.org/10.1037/0003-066X.52.11.1187

Martin, L. R., Dennehy, R., & Morgan, S. (2013). Unreliability in student evaluation of teaching questionnaires: Focus groups as an alternative approach. *Organization Management Journal, 10*(1), 66-74. http://dx.doi.org/10.1080/15416518.2013.781401

Maurer, T. W. (2006). Cognitive dissonance or revenge? Student grades and course evaluations. *Teaching of Psychology, 33*(3), 176-179. http://dx.doi.org/10.1207/s15328023top3303_4

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*(11), 1218-1225. http://dx.doi.org/10.1037/0003-066X.52.11.1218

Merritt, D. J. (2012). Bias, the brain, and student evaluations of teaching. *St. John's Law Review, 82*(1), Article 6, 235-288. http://scholarship.law.stjohns.edu/lawreview/vol82/iss1/6

Miles, P., & House, D. (2015). The tail wagging the dog: An overdue examination of student teaching evaluations. *International Journal of Higher Education, 4*(2). http://dx.doi.org/10.5430/ijhe.v4n2p116

Mitry, D. J., & Smith, D. E. (2014). Student evaluations of faculty members: A call for analytical prudence. *Journal on Excellence in College Teaching, 25*(2), 56-67. http://celt.miamioh.edu/ject/issue.php?v=25&n=2

Morley, D. D. (2012). Claims about the reliability of student evaluations of instruction: The ecological fallacy rides again. *Studies in Educational Evaluation, 38*(1), 15-20. http://dx.doi.org/10.1016/j.stueduc.2012.01.001

Nargundkar, S., & Shrikhande, M. (2012). An empirical investigation of student evaluations of instruction: The relative importance of factors. *Decision Sciences Journal of Innovative Education, 10*(1), 117-135. http://dx.doi.org/10.1111/j.1540-4609.2011.00328.x

Nargundkar, S., & Shrikhande, M. (2014). Norming of student evaluations of instruction: Impact of noninstructional factors. *Decision Sciences Journal of Innovative Education, 12*(1), 55-72. http://dx.doi.org/10.1111/dsji.12023

Otani, K., Kim, J., & Cho, J. (2012). Student evaluation of teaching (SET) in higher education: How to use SET more effectively and efficiently in public affairs education. *Journal of Public Affairs Education, 18*(3), 531-544. http://www.naspaa.org/JPAEMessenger/index_2012summer.asp

Palmer, S. (2012). Student evaluation of teaching: keeping in touch with reality. *Quality in Higher Education, 18*(3), 297-311. http://dx.doi.org/10.1080/13538322.2012.730336

Pepe, J.W., & Wang, M.C. (2012). What instructor qualities do students reward? *College Student Journal, 46*(3), 603-614. http://www.projectinnovation.biz/csj_2006.html

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile? An analytical framework for answering the question. *Quality Assurance in Education, 15*(2), 178-191. http://dx.doi.org/10.1108/09684880710748938

Rantanen, P. (2013). The number of feedbacks needed for reliable evaluation. A multilevel analysis of the reliability, stability and generalizability of students' evaluation of teaching. *Assessment & Evaluation in Higher Education, 38*(2), 224-239. http://dx.doi.org/10.1080/02602938.2011.625471

Reardon, R. C., Leierer, S. J., & Lee, D. (2014). Class meeting schedules in relation to students' grades and evaluations of teaching. *The Professional Counselor, 2*(1), 81-89. http://dx.doi.org/10.15241/rcr.2.1.81

Reisenwitz, T.H. (2015). Student evaluation of teaching: An investigation of nonresponse bias in an online context. *Journal of Marketing Education, 38*(1), 7-17.

https://doi.org/10.1177/0273475315596778
Ridley, D., & Collins, J. (2015). A suggested evaluation metric instrument for faculty members at colleges and universities. *International Journal of Education Research, 10*(1), 97-114. Retrieved from http://eds.a.ebscohost.com.login.ezproxy.library.ualberta.ca/eds/pdfviewer/pdfviewer?sid=9ff24389-d34d-43d1-83fc-6ef82bd1ad47%40sessionmgr4009&vid=2&hid=4102
Royal, K. D., & Stockdale, M. R. (2015). Are teacher course evaluations biased against faculty that teach quantitative methods courses? *International Journal of Higher Education, 4*(1), 217-224. http://dx.doi.org/10.5430/ijhe.v4n1p217
Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. *Women's Studies in Communication, 30*(1), 64-77. http://dx.doi.org/10.1080/07491409.2007.10162505
Socha, A. (2013). A hierarchical approach to students' assessment of instruction. *Assessment & Evaluation in Higher Education, 38*(1), 94-113. http://dx.doi.org/10.1080/02602938.2011.604713
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. http://dx.doi.org/10.3102/0034654313496870
Stein, S. J., Spiller, D., Terry, S., Harris, T., Deaker, L., & Kennedy, J. (2013). Tertiary teachers and student evaluations: never the twain shall meet? *Assessment & Evaluation in Higher Education, 38*(7), 892-904. http://dx.doi.org/10.1080/02602938.2013.767876
Stonebraker, R. J., & Stone, G. S. (2015). Too old to teach? The effect of age on college and university professors. *Research in Higher Education, 56*(8), 793-812. http://dx.doi.org/10.1007/s11162-015-9374-y
Stupans, I., McGuren, T., & Babey, A. M. (2016). Student evaluation of teaching: A study exploring student rating instrument free-form text comments. *Innovative Higher Education, 41*(1), 33-52. http://10.1007/s10755-015-9328-5
Uijtdehaage, S., & O'Neal, C. (2015). A curious case of the phantom professor: mindless teaching evaluations by medical students. *Medical Education, 49*(9), 928-932. http://dx.doi.org/10.1111/medu.12805
Uttl, B., White, C. A., Gonzalez, D. W. (2016). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation,* (in press, available online September 19, 2106). http://dx.doi.org/10.1016/j.stueduc.2016.08.007
Wilson, J. H., Beyer, D., & Monteiro, H. (2014). Professor age affects student ratings: Halo effect for younger teachers. *College Teaching, 62*, 20-24. http://dx.doi.org/10.1080/87567555.2013.825574
Wright, S. L., & Jenkins-Guarieri, M. A. (2012). Student evaluations of teaching: combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education, 37*(6), 683-699. http://dx.doi.org/10.1080/02602938.2011.563279
Zimmerman, B. (2008). Course evaluations - students' revenge? *University Affairs.* Retrieved

from http://www.universityaffairs.ca/opinion/in-my-opinion/course-evaluations-students-revenge/

Zumbach, J., & Funke, J. (2014). Influences of mood on academic course evaluations. *Practical Assessment, Research & Evaluation, 19*(4). http://pareonline.net/genpare.asp?wh=0&abt=19

Appendix H: Abstracts for Reviewed Literature

Click on the links to move directly to each bookmarked section. For brief summarizing points of each article, see Appendix A

Biases

Biases, Gender
Boring, Ottoboni, & Stark (2016): ratings are biased against female instructors by an amount that is large and statistically significant Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research, 2016*(1). http://dx.doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1 [Abstract, abridged] We show: SET are biased against female instructors by an amount that is large and statistically significant; The bias affects how students rate even putatively objective aspects of teaching, such as how promptly assignments are graded; The bias varies by discipline and by student gender, among other things; It is not possible to adjust for the bias, because it depends on so many factors; SET are more sensitive to students' gender bias and grade expectations than they are to teaching effectiveness; Gender biases can be large enough to cause more effective instructors to get lower SET than less effective instructors.
Centra & Gaubatz (2000): only small same-gender preferences found, particularly with females Centra, J. A., Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education, 71*(1), 17-44. http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsjsr&AN=edsjsr.10.2307.2649280&site=eds-live&scope=site [Abstract] In an attempt to determine whether male and female students rate teachers

differently depending on the gender of the teacher, we analyzed data from 741 classes in which there were at least 10 male and 10 female students. The results revealed small same gender preferences, particularly in female students rating female teachers. Teaching style rather than gender may well explain these preferences.

Gehrt, Louie, & Osland (2015): female students evaluated female lower-ranked faculty most favorably; male students evaluations were more favorable for lower ranked male faculty, but they did not degrade higher ranked female faculty

Gehrt, K., Louie, T. A., & Osland, A. (2015). Student and professor similarity: Exploring the effects of gender and relative age. *Journal of Education for Business, 90*, 1-9. http://dx.doi.org/10.1080/08832323.2014.968514

[Abstract, abridged] It was hypothesized that students would more favorably evaluate faculty who were similar in gender and in relative age (as reflected in faculty rank). As anticipated, female students evaluated female lower ranked faculty most favorably, and male higher ranked faculty least favorably. However, male students showed mixed effects. Although their evaluations were more favorable for lower ranked male faculty, they unexpectedly did not degrade higher ranked female faculty.

Huebner & Magel (2015): variances of the class average responses between male and female faculty were higher for male faculty

Huebner, L., & Magel, R. C. (2015). A gendered study of student ratings of instruction. *Open Journal of Statistics, 5,* 552-567. http://dx.doi.org/10.4236/ojs.2015.56058

[Abstract, abridged] This research tests for differences in mean class averages between male and female faculty for questions on a student rating of instruction form at one university in the Midwest. Differences in variances of class averages are also examined for male and female faculty. Tests are conducted by first considering all classes across the entire university and then classes just within the College of Science and Mathematics. The proportion of classes taught by female instructors in which the average male student rating was higher than the average female student rating was compared to the proportion of classes taught by male instructors in which the average male student rating was higher than the average female student rating.

Laube, Massoni, Sprague, & Ferber (2007): the inconsistency on the question of whether student evaluations are gendered is itself an artifact of the way that quantitative measures can mask underlying gender bias

Laube, H., Massoni, K., Sprague, J., & Ferber, A. L. (2007). The impact of gender on the evaluation of teaching: What we know and what we can do. *NWSA Journal, 19*(3), 87-104. Retrieved from http://www.jstor.org/stable/40071230

[Abstract, abridged] Scholars who have attempted to determine whether/how gender enters into students' evaluations of their teachers generally fall into two camps: those who find gender to have no (or very little) influence on evaluations, and those who find gender to affect evaluations significantly. Drawing on insights developed from sociological scholarship on gender and evaluation, we argue that the apparent inconsistency on the question of whether student evaluations are gendered is itself an artifact of the way that quantitative measures can mask underlying gender bias.

MacNell, Driscoll, & Hunt (2015): students rate males significantly higher than females

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. http://dx.doi.org/10.1007/s10755-014-9313-4

[Abstract, abridged] Although instructor gender has been shown to play an important role in influencing student ratings, the extent and nature of that role remains contested. While difficult to separate gender from teaching practices in person, it is possible to disguise an instructor's gender identity online. In our experiment, assistant instructors in an online class each operated under two different gender identities. Students rated the male identity significantly higher than the female identity, regardless of the instructor's actual gender, demonstrating gender bias.

Miles & House (2015): lower ratings for female instructors teaching larger required classes

Miles, P., & House, D. (2015). The tail wagging the dog: An overdue examination of student teaching evaluations. *International Journal of Higher Education, 4*(2). http://dx.doi.org/10.5430/ijhe.v4n2p116

[Abstract, abridged] Purpose: The purpose of this research is to examine the impact of several factors beyond the professor's control and their unique impact on Student Teaching Evaluations (STEs). The present research pulls together a substantial amount of data to statistically analyze several academic historical legends about just how vulnerable STEs are to the effects of: class size, course type, professor gender, and course grades. Design/methodology/approach: This research is utilizes over 30,000 individual student evaluations of 255 professors, spanning six semesters, during a three year time period to test six hypotheses. The final sample represents 1057 classes ranging in size between 10 and 190 students. Each hypothesis is statistically analyzed, with either analysis of variance or a Regression model. Findings: This study finds support for 5 out of 6 hypotheses. Specifically, these data suggest STEs are likely to be closest to "5" (using a 1-5 scale with 5 being highest) in small elective classes, and lowest in large required classes taught by females. As well we find support for the notion that higher expected course grades may lead to higher STEs.

Smith, Yoo, Farr, Salmon, & Miller (2007): male and female students rated female instructors more highly; effect was small but significant due to sample size

Smith, S. W., Yoo, J. H., Farr, A. C., Salmon, C. T., & Miller, V. D. (2007). The influence of student sex and instructor sex on student ratings of instructors: Results from a college of communication. *Women's Studies in Communication, 30*(1), 64-77. http://dx.doi.org/10.1080/07491409.2007.10162505

[Abstract, abridged] We posed research questions as to whether male and female students would rate male or female instructors more highly on five dimensions of student rating forms, one of which was instructor interaction. Results indicated that male and female students rated female instructors more highly on all five dimensions. The effect sizes of these results were extremely small, but significant due to the large sample size (almost 12,000). These findings suggest that administrators should not assume one sex to provide better or poorer instruction, and they should reward instructors on the basis of individual course performance rather than according to instructor sex.

Wilson, Beyer, & Monteiro (2014): lower ratings for older instructors, but more so for females than males

Wilson, J. H., Beyer, D., & Monteiro, H. (2014). Professor age affects student ratings: Halo effect for younger teachers. *College Teaching, 62*, 20-24. http://dx.doi.org/10.1080/87567555.2013.825574

[Abstract, abridged] In the present study, we examined the potential effects of professor age and gender on student perceptions of the teacher as well as their anticipated rapport in the classroom. We also asked students to rate each instructor's attractiveness based on societal beliefs about age and beauty. We expected students to rate a picture of a middle-aged female professor more negatively (and less attractive) than the younger version of the same woman. For the young versus old man offered in a photograph, we expected no age effects. Although age served as a detriment for both genders, evaluations suffered more based on aging for female than male professors.

Wright & Jenkins-Guarieri (2012): SETs appear to be valid and free from gender bias

Wright, S. L., & Jenkins-Guarieri, M. A. (2012). Student evaluations of teaching: combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education, 37*(6), 683-699. http://dx.doi.org/10.1080/02602938.2011.563279

[Abstract, abridged] Given that there is not one study summarising all these domains of research, a comprehensive overview of SETs was conducted by combining all prior meta-analyses related to SETs. Eleven meta-analyses were identified, and nine meta-analyses covering 193 studies were included in the analysis, which yielded a

small-to-medium overall weighted mean effect size (r = .26) between SETs and the variables studied. Findings suggest that SETs appear to be valid, have practical use that is largely free from gender bias and are most effective when implemented with consultation strategies.

Biases, Instructor Characteristics

Cheng (2015): tenure does not have a significant impact on student ratings of teaching performance

Cheng, D. A. (2015). Effects of professorial tenure on undergraduate ratings of teaching performance. *Education Economics, 23*(3), 338-357. http://dx.doi.org/10.1080/09645292.2013.826632

[Abstract, abridged] This study estimates the effect of professorial tenure on undergraduate ratings of learning, instructor quality, and course quality at the University of California, San Diego from Summer 2004 to Spring 2012. During this eight-year period, 120 assistant professors received tenure and 83 associate professors attained full rank. A differences-in-differences model controlling for teaching experience, study hours, response rate, and unobserved heterogeneity among terms, courses, and professors suggests that for a given professor, tenure does not have a significant impact on student ratings of teaching performance, at least in the immediate years after advancement. The results are similar for the promotion from associate to full professor.

Cho & Otani (2014): students give higher ratings for limited-term lecturers versus full-time faculty

Cho, J., & Otani, K. (2014). Differences in student evaluations of limited-term lecturers and full-time faculty. *Journal on Excellence in College Teaching, 25*(2), 5-24. http://opus.ipfw.edu/profstudies_facpubs/64

[Abstract, abridged] This study compared student evaluations of teaching (SET) for limited-term lecturers (LTLs) and full-time faculty (FTF) using a Likert-scaled survey administered to students (N = 1,410) at the end of university courses. Data were analyzed using a general linear regression model to investigate the influence of multi-dimensional evaluation items on the overall rating item (Overall, I would rate the instructor of this course as outstanding) on the SET. Results showed that students provided higher ratings for LTLs than FTF, but they value different items when rating the overall evaluation of LTLs and FTF. Some survey items (for instance, those about instructor planning and enthusiasm) influence more on the rating of the overall item for LTLs than for FTF, whereas other, multi-dimensional items (for instance, those about assessment strategies and instructor's availability) influence more on the overall rating for FTF than for LTLs.

Clayson (2013): students' first perceptions of an instructor's personality are significantly related to ratings at the end of the semester

Clayson, D. E. (2013). Initial impressions and the student evaluation of teaching. *Journal of Education for Business, 88*(1), 26-53. http://dx.doi.org/10.1080/08832323.2011.633580

[Abstract, abridged] The author looked at the initial student perceptions and conditions of a class and compared these with conditions and evaluations 16 weeks later at the end of the term. It was found that the first perceptions of the instructor and the instructor's personality were significantly related to the evaluations made at the end of the semester.

Felton, Mitchell, & Stinson (2004): students give attractively-rated professors higher quality and easiness scores

Felton, J., Mitchell, J., & Stinson, M. (2004). Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education, 29*(1), 91-108. http://dx.doi.org/10.1080/0260293032000158180

[Abstract, abridged] College students critique their professors' teaching at RateMyProfessors.com, a web page where students anonymously rate their professors on Quality, Easiness, and Sexiness. Using the self-selected data from this public forum, we examine the relations between quality, easiness, and sexiness for 3190 professors at 25 universities. For faculty with at least ten student posts, the correlation between quality and easiness is 0.61, and the correlation between quality and sexiness is 0.30. Using simple linear regression, we find that about half of the variation in quality is a function of easiness and sexiness. When grouped into sexy and non-sexy professors, the data reveal that students give sexy-rated professors higher quality and easiness scores.

Kim & MacCann (2016): students' expressed educational satisfaction was related to perceptions of instructor personality

Kim, L. E., MacCann, C. (2016). What is students' ideal university instructor personality? An investigation of absolute and relative personality preferences. *Personality and Individual Differences, 102*, 190-203. http://dx.doi.org/10.1016/j.paid.2016.06.068

[Abstract, abridged] The current two studies investigate students' descriptions of "ideal" instructor personality using the Five-Factor Model of personality. Both absolute personality preferences (certain traits are universally desired) and relative personality preferences (certain traits are desired relative to students' own level of the trait) are examined among 137 first year mathematics students (Study 1) and 378 first year psychology students (Study 2). Students provided Big Five personality ratings for themselves, their actual instructor, and their ideal instructor. Supporting the absolute preference hypothesis, students rated their ideal instructor as having significantly higher levels than both themselves and the general population on all five personality domains (except for openness in Study 1), with particularly large effect sizes for emotional stability and conscientiousness. Supporting the relative preference hypothesis, students also rated their ideal instructor as having a similar Big Five profile to themselves. Moreover, if their actual instructor's personality was similar to their ideal instructor's personality, students showed greater educational satisfaction (but not higher performance self-efficacy nor academic achievement).

Stonebraker & Stone (2015): age has a negative impact on student ratings of faculty members; begins around mid-forties; offset by attractiveness

Stonebraker, R. J., & Stone, G. S. (2015). Too old to teach? The effect of age on college and university professors. *Research in Higher Education, 56*(8), 793-812. http://dx.doi.org/10.1007/s11162-015-9374-y

[Abstract, abridged] Using data from the RateMyProfessors.com website for a large sample of instructors in a broad cross-section of colleges and universities, we find that age does affect teaching effectiveness, at least as perceived by students. Age has a negative impact on student ratings of faculty members that is robust across genders, groups of academic disciplines and types of institutions. However, the effect does not begin until faculty members reach their mid-forties and does not seem to increase even when they reach the former retirement ages of 65 or 70. Moreover, the quantitative impact of age on student ratings is small and can be offset by other factors, especially the physical appearance of professors and how easy students consider them to be. When we restrict our sample to those professors deemed hot by student raters, the effect of age disappears completely.

Wilson, Beyer, & Monteiro (2014): lower ratings for older instructors, but more so for females than males

Wilson, J. H., Beyer, D., & Monteiro, H. (2014). Professor age affects student ratings: Halo effect for younger teachers. *College Teaching, 62*, 20-24. http://dx.doi.org/10.1080/87567555.2013.825574

[Abstract, abridged] In the present study, we examined the potential effects of professor age and gender on student perceptions of the teacher as well as their anticipated rapport in the classroom. We also asked students to rate each instructor's attractiveness based on societal beliefs about age and beauty. We expected students to rate a picture of a middle-aged female professor more negatively (and less attractive) than the younger version of the same woman. For the young versus old man offered in a photograph, we expected no age effects. Although age served as a detriment for both genders, evaluations suffered more based on aging for female than male professors.

Biases, Correlation Between Grades and Ratings

Backer (2012): some students punish academics for failing grades with low ratings

Backer, E. (2012). Burnt at the student evaluation stake – the penalty for failing students. *E-Journal of Business Education & Scholarship of Teaching, 6*(1), 1-13. Retrieved from http://www.ejbest.org/upload/eJBEST_Backer_2012_1.pdf

[Abstract, abridged] Despite the wealth of research in the area of SETs, little has been done

to examine student and academic perceptions of SETs. This research examined student (n=235) and academic (n=49) perceptions concerning SETs at one Australian regional university. Almost one-third of respondents felt that some students punish academics for failing their work by giving the lecturer low scores on the SET form. Thus, academics can essentially be burnt at the student evaluation stake as punishment for failing students.

Blackhart, Peruche, DeWall, & Joiner (2006): higher ratings given to instructors who give higher grades, and also to graduate teaching assistant rank

Blackhart, G. C., Peruche, B .M., DeWall, C. N., & Joiner, T. E., Jr. (2006). Faculty forum: Factors influencing teaching evaluations in higher education. *Teaching of Psychology, 33*(1), 37-39. http://dx.doi.org/10.1207/s15328023top3301_9

[Abstract, abridged] Past research indicates several factors influencing teaching evaluation ratings instructors receive. We analyzed teaching evaluations from psychology courses during fall and spring semesters of 2003– 2004 to determine if class size, class level, instructor gender, number of publications (faculty instructors), average grade given by the instructor, and instructor rank predicted teaching evaluation ratings. Entering predictor variables into a multiple regression analysis concurrently, results indicated that only average grade given and instructor rank significantly predicted instructor ratings. Specifically, higher average grades given by the instructor predicted higher ratings, and graduate teaching assistants received higher overall ratings than faculty instructors.

Boring, Ottoboni, & Stark (2016): ratings are more sensitive to students' grade expectations than they are to teaching effectiveness

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

[Abstract, abridged] We show: SET are biased against female instructors by an amount that is large and statistically significant; The bias affects how students rate even putatively objective aspects of teaching, such as how promptly assignments are graded; The bias varies by discipline and by student gender, among other things; It is not possible to adjust for the bias, because it depends on so many factors; SET are more sensitive to students' gender bias and grade expectations than they are to teaching effectiveness; Gender biases can be large enough to cause more effective instructors to get lower SET than less effective instructors.

Centra (2003): expected grades generally do not affect student evaluations

Centra, J.A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*(5), 495-518. http://www.jstor.org.login.ezproxy.library.ualberta.ca/stable/40197319

[Abstract, abridged] This study investigated whether mean expected grades and the level of difficult/workload in courses, as reported by students, unduly influence student ratings

instruction. Over 50,000 college courses were analyzed. After controlling for learning outcomes, expected grades generally did not affect student evaluations. In fact, contrary to what some faculty think, courses in natural sciences with expected grades of A were rated lower, not higher. Courses were rated lower when they were rated as either difficult or too elementary. Courses rated at the "just right" level received the highest evaluations.

Cho, Baek, & Cho (2015): students with better grades than their expected grades provide a psychological "gift" to their teachers by giving higher ratings

Cho, D., Baek, W., & Cho, J. (2015). Why do good performing students highly rate their instructors? Evidence from a natural experiment. *Economics of Education Review, 49*, 172-179. http://dx.doi.org/10.1016/j.econedurev.2015.10.001

[Abstract, abridged] This article analyzes the behavior of students in a college classroom with regard to their evaluation of teacher performance. As some students are randomly able to see their grades prior to the evaluation, the "natural" experiment provides a unique opportunity for testing the hypothesis as to whether there exists a possibility of a hedonic (implicit) exchange between the students' grades and teaching evaluations. Students with good grades tend to highly rate the teaching quality of their instructors, in comparison with those who receive relatively poor grades. This study finds that students with better grades than their expected grades provide a psychological "gift" to their teachers by giving a higher teacher evaluation, whereas it is the opposite with those students receiving lower grades than their expectation.

Greenwald & Gillmore (1997): the grades-ratings correlation is due to an unwanted influence of instructors' grading leniency; there are 5 theories of the grades-ratings correlation

Greenwald, A. G., Gillmore, G. M. (1997). Grade leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209-1217. http://dx.doi.org/10.1037/0003-066X.52.11.1209

[Abstract] It is well established that students' evaluative ratings of instruction correlate positively with expected course grades. The authors identify 4 additional data patterns that, collectively, discriminate among 5 theories of the grades-ratings correlation. The presence of all 4 of these markers in student ratings data (obtained at University of Washington) was most consistent with the theory that the grades-ratings correlation is due to an unwanted influence of instructors' grading leniency on ratings. This conclusion justifies use of a statistical correction – illustrated here with actual ratings data – to remove the unwanted inflation of ratings produced by lenient grading. Additional research can profitably seek other inappropriate influences on ratings to identify more opportunities for validity-enhancing adjustments.

Gump (2007): questions the validity of research done on the leniency hypothesis

Gump, S.E. (2007). Student evaluations of teaching effectiveness and the leniency hypothesis: A literature review. *Education Research Quarterly, 30*(3), 55-68. Retrieved from http://eric.ed.gov.login.ezproxy.library.ualberta.ca/?id=EJ787711

[Abstract, abridged] This review presents an overview of selected articles on the leniency hypothesis: the idea that students give higher evaluations to instructors who grade more leniently. In this diverse literature, research methods and aims have frequently affected the outcomes and conclusions, since SETs are typically context-specific instruments whose results, in isolated instances, do not generalize well. Thus this review questions the very generalizability of the massive and often contradictory SET-related literature on the leniency hypothesis and argues that future research must be designed and carried out in light of the implicit problems existing in the majority of earlier studies.

Maurer (2006): cognitive dissonance may be a theory to explain the grades-ratings correlation

Maurer, T. W. (2006). Cognitive dissonance or revenge? Student grades and course evaluations. *Teaching of Psychology, 33*(3), 176-179. http://dx.doi.org/10.1207/s15328023top3303_4

[Abstract] I tested 2 competing theories to explain the connection between students' expected grades and ratings of instructors: cognitive dissonance and revenge. Cognitive dissonance theory holds that students who expect poor grades rate instructors poorly to minimize ego threat whereas the revenge theory holds that students rate instructors poorly in an attempt to punish them. I tested both theories via an experimental manipulation of the perceived ability to punish instructors through course evaluations. Results indicated that student ratings appear unrelated to the ability to punish instructors, thus supporting cognitive dissonance theory. Alternative interpretations of the data suggest further research is warranted.

Miles & House (2015): higher expected grades may lead to higher ratings

Miles, P., & House, D. (2015). The tail wagging the dog: An overdue examination of student teaching evaluations. *International Journal of Higher Education, 4*(2). http://dx.doi.org/10.5430/ijhe.v4n2p116

[Abstract, abridged] Purpose: The purpose of this research is to examine the impact of several factors beyond the professor's control and their unique impact on Student Teaching Evaluations (STEs). The present research pulls together a substantial amount of data to statistically analyze several academic historical legends about just how vulnerable STEs are to the effects of: class size, course type, professor gender, and course grades. Design/methodology/approach: This research is utilizes over 30,000 individual student evaluations of 255 professors, spanning six semesters, during a three year time period to test six hypotheses. The final sample represents 1057 classes ranging in size between 10 and 190 students. Each hypothesis is statistically analyzed, with either analysis of variance or a Regression model. Findings: This study finds support for 5 out of 6 hypotheses. Specifically,

these data suggest STEs are likely to be closest to "5" (using a 1-5 scale with 5 being highest) in small elective classes, and lowest in large required classes taught by females. As well we find support for the notion that higher expected course grades may lead to higher STEs.

Biases, Nonresponse

Kuwaiti, AlQuraan, & Subbarayalu (2016): ratings are affected by class size and response rate

Kuwaiti, A. A., AlQuraan, M., & Subbarayalu, A. V. (2016). Understanding the effect of response rate and class size interaction on students evaluation of teaching in a higher education. *Educational Assessment & Evaluation, 3*, https://doi.org/10.1080/2331186X.2016.1204082

[Abstract, abridged] This study aims to investigate the interaction between response rate and class size and its effects on students' evaluation of instructors and the courses offered at a higher education Institution in Saudi Arabia. It is observed that when the class size is at the medium level, the ratings of instructors and courses increase as the response rate increases. On the contrary; when the class size is small, a high response rate is required for the evaluation of instructors and at least medium response rate is required for evaluation of courses. The study suggests that the interaction between response rate and class size is an important factor that needs to be taken into account while interpreting the students' evaluation of instructors and courses.

Macfadyen, Dawson, Prest, & Gasevic (2016): much bias based on who is completing the surveys

Macfadyen, L. P., Dawson, S., Prest, S., & Gasevic, D. (2016). Whose feedback? A multilevel analysis of student completion of end-of-term teaching evaluations. *Assessment & Evaluation in Higher Education, 41*(6), 821-839. http://dx.doi.org/10.1080/02602938.2015.1044421

[Abstract, abridged] While much research has examined the validity of SETs for measuring teaching quality, few studies have investigated the factors that influence student participation in the SET process. This study aimed to address this deficit through the analysis of an SET respondent pool at a large Canadian research-intensive university. The findings were largely consistent with available research (showing influence of student gender, age, specialisation area and final grade on SET completion). However, the study also identified additional influential course-specific factors such as term of study, course year level and course type as statistically significant. Collectively, such findings point to substantively significant patterns of bias in the characteristics of the respondent pool.

Reisenwitz (2015): there are significant differences between those who complete online student evaluations and those who do not

Reisenwitz, T.H. (2015). Student evaluation of teaching: An investigation of nonresponse bias

in an online context. *Journal of Marketing Education, 38*(1), 7-17. https://doi.org/10.1177/0273475315596778

[Abstract, abridged] This study examines nonresponse bias in online student evaluations of instruction, that is, the differences between those students who complete online evaluations and those who decide not to complete them. It builds on the work of Estelami that revealed a response bias based on the timing in which the evaluations were completed, that is, differences in early evaluations versus later evaluations. In contrast, this study examines the demographic variables that have contributed to nonresponse bias in online student evaluations, namely gender, grade point average, and ethnicity. It also examines multiple psychographic variables that may contribute to nonresponse bias: time poverty, complaining behavior, and technology savviness. This study found that there are significant differences between those who complete online student evaluations and those who do not.

Biases, Non-instructional

Kuwaiti, AlQuraan, & Subbarayalu (2016): ratings are affected by class size and response rate

Kuwaiti, A. A., AlQuraan, M., & Subbarayalu, A. V. (2016). Understanding the effect of response rate and class size interaction on students evaluation of teaching in a higher education. *Educational Assessment & Evaluation, 3*, https://doi.org/10.1080/2331186X.2016.1204082

[Abstract, abridged] This study aims to investigate the interaction between response rate and class size and its effects on students' evaluation of instructors and the courses offered at a higher education Institution in Saudi Arabia. It is observed that when the class size is at the medium level, the ratings of instructors and courses increase as the response rate increases. On the contrary; when the class size is small, a high response rate is required for the evaluation of instructors and at least medium response rate is required for evaluation of courses. The study suggests that the interaction between response rate and class size is an important factor that needs to be taken into account while interpreting the students' evaluation of instructors and courses.

Nargundkar & Shrikhande (2014): combined impact of all the noninstructional factors studied is statistically significant

Nargundkar, S., & Shrikhande, M. (2014). Norming of student evaluations of instruction: Impact of noninstructional factors. *Decision Sciences Journal of Innovative Education, 12*(1), 55-72. http://dx.doi.org/10.1111/dsji.12023

[Abstract, abridged] Student Evaluations of Instruction (SEIs) from about 6,000 sections over 4 years representing over 100,000 students at the college of business at a large public university are analyzed, to study the impact of noninstructional factors on student ratings. Administrative factors like semester, time of day, location, and instructor attributes like gender

and rank are studied. The combined impact of all the noninstructional factors studied is statistically significant. Our study has practical implications for administrators who use SEIs to evaluate faculty performance. SEI scores reflect some inherent biases due to noninstructional factors. Appropriate norming procedures can compensate for such biases, ensuring fair evaluations.

Reardon, Leierer, & Lee (2014): class schedule does not affect ratings

Reardon, R. C., Leierer, S. J., & Lee, D. (2014). Class meeting schedules in relation to students' grades and evaluations of teaching. *The Professional Counselor, 2*(1), 81-89. http://dx.doi.org/10.15241/rcr.2.1.81

[Abstract, abridged] A six-year retrospective study of a university career course evaluated the effect of four different class schedule formats on students' earned grades, expected grades and evaluations of teaching. Some formats exhibited significant differences in earned and expected grades, but significant differences were not observed in student evaluations of instruction.

Royal & Stockdale (2015): students give lower ratings to instructors of quantitative methods subjects

Royal, K. D., & Stockdale, M. R. (2015). Are teacher course evaluations biased against faculty that teach quantitative methods courses? *International Journal of Higher Education, 4*(1), 217-224. http://dx.doi.org/10.5430/ijhe.v4n1p217

[Abstract, abridged] The present study investigated graduate students' responses to teacher/course evaluations (TCE) to determine if students' responses were inherently biased against faculty who teach quantitative methods courses. Item response theory (IRT) and Differential Item Functioning (DIF) techniques were utilized for data analysis. Results indicate students in non-methods courses preferred the structure of quantitative courses, but tend to be more critical of quantitative instructors.

Biases, Other

Blackhart, Peruche, DeWall, & Joiner (2006): varying results for investigation if class size, class level, instructor gender, number of publications (faculty instructors), average grade given by the instructor, and instructor rank predicted teaching evaluation ratings

Blackhart, G. C., Peruche, B. M., DeWall, C. N., & Joiner, T. E., Jr. (2006). Faculty forum: Factors influencing teaching evaluations in higher education. *Teaching of Psychology, 33*(1), 37-39. http://dx.doi.org/10.1207/s15328023top3301_9

[Abstract, abridged] Past research indicates several factors influencing teaching evaluation ratings instructors receive. We analyzed teaching evaluations from psychology courses during

fall and spring semesters of 2003-2004 to determine if class size, class level, instructor gender, number of publications (faculty instructors), average grade given by the instructor, and instructor rank predicted teaching evaluation ratings. Entering predictor variables into a multiple regression analysis concurrently, results indicated that only average grade given and instructor rank significantly predicted instructor ratings. Specifically, higher average grades given by the instructor predicted higher ratings, and graduate teaching assistants received higher overall ratings than faculty instructors.

Keeley, English, Irons, & Henslee (2013): found halo and ceiling/floor effects to be present and persistent

Keeley, J. W., English, T., Irons, J., & Henslee, A. M. (2013). Investigating halo and ceiling effects in student evaluations of instruction. *Educational and Psychological Measurement, 73*(3), 440-457. http://dx.doi.org/10.1177/0013164412475300

[Abstract, abbreviated, and other article text] Many measurement biases affect student evaluations of instruction (SEIs). However, two have been relatively understudied: halo effects and ceiling/floor effects. This study examined these effects in two ways. Both biases were robust and remained despite characteristics of the measure designed to combat them.

“halo effects occur when a rater’s opinion about one aspect of the teacher influences the remainder of that person’s ratings”

“Ceiling and floor effects (also referred to as maximizing and minimizing effects) occur when a scale does not have a sufficient range to produce meaningful variability at the upper or lower ends of possible scores.”

Marsh & Roche (1997): evaluations are valid and unaffected by hypothesized biases

Marsh, H. W., & Roche, L. A. (1997). Making students’ evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187-1197. http://dx.doi.org/10.1037/0003-066X.52.11.1187

[Abstract, abridged] This article reviews research indicating that, under appropriate conditions, students' evaluations of teaching (SETs) are (a) multidimensional; (b) reliable and stable; (c) primarily a function of the instructor who teaches a course rather than the course that is taught; (d) relatively valid against a variety of indicators of effective teaching; (e) relatively unaffected by a variety of variables hypothesized as potential biases (e.g., grading leniency, class size, workload, prior subject interest); and (f) useful in improving teaching effectiveness when SETS are coupled with appropriate consultation. The authors recommend rejecting a narrow criterion-related approach to validity and adopting a broad construct-validation approach, recognizing that effective teaching and SETs that reflect teaching effectiveness are multidimensional; no single criterion of effective teaching is sufficient; and tentative interpretations of relations with validity criteria and potential biases should be evaluated critically in different contexts, in relation to multiple criteria of effective teaching, theory, and existing knowledge.

Merritt (2012): covers biases in general, including race minority

Merritt, D. J. (2012). Bias, the brain, and student evaluations of teaching. *St. John's Law Review, 82*(1), Article 6, 235-288. http://scholarship.law.stjohns.edu/lawreview/vol82/iss1/6

[It seems that a 2008 version of this article was used in the UA report, but the version now online is 2012. No abstract.]

Pounder (2007): identifies and organizes factors influencing SET scores; literature review

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile? An analytical framework for answering the question. *Quality Assurance in Education, 15*(2), 178-191. http://dx.doi.org/10.1108/09684880710748938

[Abstract, abridged] Identifies student related, course related and teacher related aspects of research on teaching evaluations. Factors commonly addressed within these aspects are also identified. On the basis of a comprehensive survey of the literature, this paper identifies and discusses the central factors influencing SET scores. These factors are then presented in a comprehensible table that can be used as a reference point for researchers and practitioners wishing to examine the effectiveness of the SET system.

Zumback & Funke (2014): students' mood affects ratings

Zumbach, J., & Funke, J. (2014). Influences of mood on academic course evaluations. *Practical Assessment, Research & Evaluation, 19*(4). http://pareonline.net/genpare.asp?wh=0&abt=19

[Abstract, abridged] In two subsequent experiments, the influence of mood on academic course evaluation is examined. By means of facial feedback, either a positive or a negative mood was induced while students were completing a course evaluation questionnaire during lectures. Results from both studies reveal that a positive mood leads to better ratings of different dimensions of lecture quality. While in Study 1 (N=109) mood was not directly controlled, Study 2 (N=64) replicates the findings of the prior study and reveals direct influences of positive and negative mood on academic course evaluation.

Validity

Al-Eidan, Baig, Magzoub, & Omair (2016): the faculty evaluation tool was found to be reliable, but validity has to be interpreted with caution because of low response

Al-Eidan, F., Baig, L. A., Magzoub, M., & Omair, A. (2016). Reliability and validity of the faculty evaluation instrument used at King Saud bin Abdulaziz University for Health Sciences: Results from the haematology course. *The Journal of the Pakistan Medical Association, 66*(4), 453-457. http://www.jpma.org.pk/full_article_text.php?article_id=7711

[Abstract, abridged] Objectives: To assess reliability and validity of evaluation tool using Haematology course as an example. Results: Of the 116 subjects in the study, 80(69%) were males and 36(31%) were females. Reliability of the questionnaire was Cronbach's alpha 0.91. Factor analysis yielded a logically coherent 7 factor solution that explained 75% of the variation in the data. The factors were group dynamics in problem-based learning (alpha0.92), block administration (alpha 0.89), quality of objective structured clinical examination (alpha 0.86), block coordination (alpha 0.81), structure of problem-based learning (alpha 0.84), quality of written exam (alpha 0.91), and difficulty of exams (alpha0.41). Female students' opinion on depth of analysis and critical thinking was significantly higher than that of the males (p=0.03). Conclusion: The faculty evaluation tool used was found to be reliable, but its validity, as assessed through factor analysis, has to be interpreted with caution as the responders were less than the minimum required for factor analysis.

Bedggood & Donovan (2012): student satisfaction does not equal teaching quality; both student satisfaction and student learning are relevant measures

Bedggood, R. E., & Donovan, J. D. (2012). University performance evaluations: What are we really measuring? *Studies in Higher Education, 37*(7), 825-842. http://dx.doi.org/10.1080/03075079.2010.549221

[Abstract, abridged] Despite the criticisms surrounding whether measures associated with these surveys are indeed valid, university managers continue to utilise them in key decision making. However, some argue that universities are misdirected in measuring satisfaction as a proxy for teaching quality, possibly subverting the potentially conflicting objective of student learning. Even so, both student satisfaction and student learning can be relevant performance measures. Accordingly, we have developed two robust measures of these constructs. We argue that student learning can be measured and used to provide formative feedback for improving teaching effectiveness. Alternatively, student satisfaction can be appropriate for determining whether students are ‘enjoying’ their studies, and likewise offers distinct benefits to university managers measuring performance outcomes.

Brown, Wood, Ogden, & Maltby (2014): students’ satisfaction rating is context dependent; objective quality and subjective satisfaction are different things and should be assessed accordingly

Brown, G. D. A., Wood, A. M., Ogden, R. S., & Maltby, J. (2014). Do student evaluations of university reflect inaccurate beliefs or actual experience? A relative rank model. *Journal of Behavioral Decision Making, 28*, 14-26. http://dx.doi.org/10.1002/bdm.1827

[Abstract] It was shown that student satisfaction ratings are influenced by context in ways that have important theoretical and practical implications. Using questions from the UK’s National Student Survey, the study examined whether and how students’ expressed satisfaction with issues such as feedback promptness and instructor enthusiasm depends on the context of comparison (such as possibly inaccurate beliefs about the feedback promptness or enthusiasm experienced at other universities) that is evoked. Experiment 1 found strong effects of experimentally provided comparison context—for example, satisfaction with a given feedback time depended on the time’s relative position within a context. Experiment 2 used a

novel distribution-elicitation methodology to determine the prior beliefs of individual students about what happens in universities other than their own. It found that these beliefs vary widely and that students' satisfaction was predicted by how they believed their experience ranked within the distribution of others' experiences. A third study found that relative judgment principles also predicted students' intention to complain. An extended model was developed to show that purely rank-based principles of judgment can account for findings previously attributed to range effects. It was concluded that satisfaction ratings and quality of provision are different quantities, particularly when the implicit context of comparison includes beliefs about provision at other universities. Quality and satisfaction should be assessed separately, with objective measures (such as actual times to feedback), rather than subjective ratings (such as satisfaction with feedback promptness), being used to measure quality wherever practicable.

Chen & Hoshower (2003): student motivation to participate in SET affects ratings

Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: an assessment of student perception and motivation. *Assessment & Evaluation in Higher Education, 28*(1), 71-88. http://dx.doi.org/10.1080/0260293032000033071

[Abstract, abridged] Very few studies have looked into students' perception of the teaching evaluation system and their motivation to participate. This study employs expectancy theory to evaluate some key factors that motivate students to participate in the teaching evaluation process. The results show that students generally consider an improvement in teaching to be the most attractive outcome of a teaching evaluation system. The second most attractive outcome was using teaching evaluations to improve course content and format. Using teaching evaluations for a professor's tenure, promotion and salary rise decisions and making the results of evaluations available for students' decisions on course and instructor selection were less important from the students' standpoint. Students' motivation to participate in teaching evaluations is also impacted significantly by their expectation that they will be able to provide meaningful feedback.

Chonko, Tanner, & Davis (2002): students focus more on qualities that make a course appealing, not learning

Chonko, L. B., Tanner, J. F., & Davis, R. (2002). What are they thinking? Students' expectations and self-assessments. *Journal of Education for Business, 77*(5), 271-281. Retrieved from http://login.ezproxy.library.ualberta.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=7214031&site=eds-live&scope=site

[Abstract] Student teacher evaluations have been the subject of a great deal of research. In this study, the authors surveyed 750 freshmen in an Introduction to Business class. The authors found that students' actual perceptions often diverged from what they were assessing on teaching evaluations and that their expectations of the teacher and the class, as well as their self-assessments, were very related to how students rate classes and teachers. The authors suggest that caution should be exercised in the use of student evaluations.

Cohen (1981): student ratings are a valid measure of teaching effectiveness; this is the meta-analysis targeted by Uttl et al., 2016

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*(3), 281-309.

[Abstract, abridged] The data for the meta-analysis came from 41 independent validity studies reporting on 68 separate multisection courses relating student ratings to student achievement. A hierarchical multiple regression analysis showed that rating/achievement correlations were larger for full-time faculty when students knew their final grades before rating instructors and when an external evaluator graded students' achievement tests. The results of the meta-analysis provide strong support for the validity of student ratings as measures of teaching effectiveness.

d'Apollonia & Abrami (1997): student ratings are moderately valid; however, they are affected by administrative, instructor, and course characteristics

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist, 52*(11), 1198-1208. http://dx.doi.org/10.1037/0003-066X.52.11.1198

[Abstract, abridged] Many colleges and universities have adopted the use of student ratings of instruction as one (often the most influential) measure of instructional effectiveness. In this article, the authors present evidence that although effective instruction may be multidimensional, student ratings of instruction measure general instructional skill, which is a composite of three subskills: delivering instruction, facilitating interactions, and evaluating student learning.The authors subsequently report the results of a meta-analysis of the multisection validity studies that indicate that student ratings are moderately valid; however, administrative, instructor, and course characteristics influence student ratings of instruction.

Dodeen (2013): validity of SET is questionable

Dodeen, H. (2013). Validity, reliability, and potential bias of short forms of students' evaluation of teaching: The case of UAE University. *Educational Assessment, 18*(4), 235-250. http://dx.doi.org/10.1080/10627197.2013.846670

[Abstract, abridged] Students' opinions continue to be a significant factor in the evaluation of teaching in higher education institutions. The purpose of this study was to psychometrically assess short students evaluation of teaching (SET) forms using the UAE University form as a model. The study evaluated the form validity, reliability, the overall question, and potential bias with respect to gender, college, grade point average, expected grade, and class size. A total of 3,661 students participated in this study in different random samples. Results indicated that the short SET form lacked content validity and could not identify key dimensions of evaluating teaching effectiveness. The form showed stability over time and acceptable internal reliability. Results indicated also that there was a potential bias due to college, expected grade, and class size, but there was no relationship between grade point average and students' ratings. It was concluded that short SET forms do not cover all domain content

and unable to provide teachers with enough information for the improvement of teaching.

Dolmans, Janssen-Noordman, & Wolfhagen (2006): students can distinguish excellent and poor teaching quality

Dolmans, D. M., Janssen-Noordman, A., & Wolfhagen, H. P. (2006). Can students differentiate between PBL tutors with different tutoring deficiencies? Medical Teacher, 28(6), 156-161. doi: 10.1080/01421590600776545

[Abstract, abridged] Although everyone will agree that students are able to distinguish between poor and excellent tutors, one can question whether students are also able to differentiate between tutors with different tutoring deficiencies—tutors who perform badly on a specific key aspect of their performance. The aim of this study was to investigate to what degree students are able to differentiate between tutors with different tutoring deficiencies, how effective tutors are with different deficiencies and what kind of tips students give for improvement of a tutor's behaviour. The results of this study demonstrate that students are not only able to distinguish between poor and excellent tutors, but are also able to diagnose tutors with different tutoring deficiencies and are able to provide tutors with specific feedback to improve their performance.

Ginns, Prosser, & Barrie (2007): the SET tool studied supports quality assurance and improvement processes at the university

Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: the perspective of currently enrolled students. *Studies in Higher Education, 32*(5), 603-615. http://dx.doi.org/10.1080/03075070701573773

[Abstract, abridged] The psychometric properties of a version of the Course Experience Questionnaire revised for students currently enrolled at the University of Sydney, the Student Course Experience Questionnaire (SCEQ), were assessed, gathering students' perceptions on a number of scales, including Good Teaching, Clear Goals and Standards, Appropriate Assessment, Appropriate Workload, and an outcome scale measuring Generic Skills development. Confirmatory factor analyses supported the hypothesised factor structure, and estimates of inter-rater agreement on SCEQ scales indicated student ratings of degrees can be meaningfully aggregated up to the faculty level. Derived from a substantial research base, linking the student experience to approaches to study and learning outcomes, its goal is to support both quality assurance and improvement processes within the university, at both the degree level and faculty level. The analyses described above indicate that the SCEQ is appropriate for these purposes.

Grammatikopoulos, Linardakis, Gregoriadis, & Oikonomidis (2015): provides evidence of a valid SET instrument; evaluating test validity is a continuous process, not a one-time event

Grammatikopoulos, V., Linardakis, M., Gregoriadis, A., & Oikonomidis, V. (2015). Assessing the students' evaluations of educational quality (SEEQ) questionnaire in Greek higher education. *Higher Education, 70*(3), 395-408. http://dx.doi.org/10.1007/s10734-014-9837-7

[Abstract, abridged] The aim of the current study was to provide a valid and reliable instrument for the evaluation of the teaching effectiveness in the Greek higher education system. Other objectives of the study were (a) the examination of the dimensionality and the higher-order structure of the Greek version of Students' Evaluation of Educational Quality (SEEQ) questionnaire, and (b) the investigation of the effects of several background variables on students' evaluations of teaching (SET) scores provided by the Greek version of SEEQ. A total of 1,264 students participated by filling in the questionnaires administered to them. The results showed solid evidence of the applicability of the Greek version of SEEQ, by confirming the factor structure of the instrument and reassuring the multidimensionality of the teaching effectiveness construct. Additionally, the effects of several background variables on teaching effectiveness further supported the validity of SET scores.

Grayson (2015): questions student's ability to give accurate ratings

Grayson, J. P. (2015). Repeated low teaching evaluations: A form of habitual behavior? *Canadian Journal of Higher Education, 45*(4), 298-321. http://journals.sfu.ca/cjhe/index.php/cjhe/article/view/184404

[Abstract, abridged] In this article, comparisons were made between first- and third-year collective evaluations of professors' performance at the University of British Columbia, York University, and McGill University. Overall, it was found that students who provided low evaluations in their first year were also likely to do so in their third year. Given that over the course of their studies, students likely would have been exposed to a range of different behaviours on the part of their professors, it is argued that the propensity of a large number of students to give consistently low evaluations was a form of "habitual behaviour.

Greenwald (1997): student rating measures have validity concerns

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182-1186. http://dx.doi.org/10.1037/0003-066X.52.11.1182

[Abstract] The validity of student rating measures of instructional quality was severely questioned in the 1970s. By the early 1980s, however, most expert opinion viewed student rating measures as valid and as worthy of widespread use. In retrospect, older discriminant-validity concerns were not so much resolved as they were displaced from research attention by accumulating evidence for convergent validity. This article introduces a Current Issues section that gives new attention to validity concerns associated with student ratings. The section's 4 articles deal, respectively, with (a) conceptual structure (are student ratings unidimensional or multidimensional?), (b) convergent validity (how well do ratings correlate with other indicators of effective teaching?), (c) discriminant validity (are ratings influenced by factors other than teaching effectiveness?), and (d) consequential validity (are ratings used effectively in personnel development and evaluation?). Although all 4 articles favor the use of ratings, they disagree on controversial points associated with interpretation and use of ratings data.

Khong (2014): SET is a valid instrument in evaluating teaching effectiveness

Khong, T. L. (2014). The validity and reliability of the student evaluation of teaching: A case in a private higher educational institution in Malaysia. *International Journal for Innovation Education and Research, 2*(9), 57-63. http://www.ijier.net/index.php/ijier/article/view/317

[Abstract, abridged] Most universities are using the Student Evaluation of Teaching (SET) as an instrument for students to assess a lecturer's teaching performance. It is an essential instrument to reflect the feedback in enhancing the quality of teaching and learning. The purpose of this paper is to examine the validity and reliability of the SET as a valid instrument in evaluating teaching effectiveness in a private higher education institution in Malaysia. Exploratory Factor Analysis and Confirmatory Factor Analysis have validated all 10 items of SET whereby all items indicated high reliability and internal consistency.

The conclusion of this study showed that the SET is a valid instrument in evaluating teaching effectiveness.

Lama, Arias, Mendoza, & Manahan (2015): lack of student diligence when rating instructors raises validity concerns

Lama, T., Arias, P., Mendoza, K. & Manahan, J. (2015). Student evaluation of teaching surveys: do students provide accurate and reliable information? *e-Journal of Social & Behavioural Research in Business, 6*(1), 30-39. http://www.ejsbrb.org/a.php?/content/issue/10

[Abstract, abridged] This paper explores patterns of students' response behaviour of international students studying in an Australian university when filling out student surveys evaluating lecturers and courses. The study focuses on whether information obtained through the survey process can be relied upon to make management decisions. The results of the study seem to suggest a reasonable level of diligence is lacking on the students' part in answering the surveys, raising a concern about the reliability of information. This tendency seems to be prevalent among all students irrespective of their gender and nationality.

Marsh & Roche (1997): evaluations are relatively valid and unaffected by hypothesized biases

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187-1197. http://dx.doi.org/10.1037/0003-066X.52.11.1187

[Abstract, abridged] This article reviews research indicating that, under appropriate conditions, students' evaluations of teaching (SETs) are (a) multidimensional; (b) reliable and stable; (c) primarily a function of the instructor who teaches a course rather than the course that is taught; (d) relatively valid against a variety of indicators of effective teaching; (e) relatively unaffected by a variety of variables hypothesized as potential biases (e.g., grading leniency, class size, workload, prior subject interest); and (f) useful in improving teaching effectiveness when SETS are coupled with appropriate consultation. The authors recommend rejecting a narrow criterion-related approach to validity and adopting a broad construct-validation approach, recognizing that effective teaching and SETs that reflect teaching effectiveness are multidimensional; no single criterion of effective teaching is sufficient; and tentative

interpretations of relations with validity criteria and potential biases should be evaluated critically in different contexts, in relation to multiple criteria of effective teaching, theory, and existing knowledge.

Martin, Dennehy, & Morgan (2013): validity of SET is questioned; student focus groups suggested as an alternative

Martin, L. R., Dennehy, R., & Morgan, S. (2013). Unreliability in student evaluation of teaching questionnaires: Focus groups as an alternative approach. *Organization Management Journal, 10*(1), 66-74. http://dx.doi.org/10.1080/15416518.2013.781401

[Abstract, abridged] Research on the validity and reliability of SETs is vast, though riddled with inconsistencies. The many "myths" of SETs are investigated and the incongruities are demonstrated. We hypothesize that the discrepancies in empirical studies come from misunderstanding and inappropriate actions by students. To address the complexity inherent in these problems, we suggest the use of focus groups as an alternative approach or complement to the standard SETs. A recommended format and guidelines for running classroom focus groups are provided. Institutional constraints and implementation concerns are addressed as well. This article lays the foundation for implementing a change in student assessment of teaching by proposing a method to compensate for bias in SETs, using focus groups as an evaluation tool, either as a stand-alone process or as a supplement to current methods.

McKeachie (1997): student ratings are valid but affected by contextual variables such as grading leniency

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*(11), 1218-1225. http://dx.doi.org/10.1037/0003-066X.52.11.1218

[Abstract, abridged] In this article, the author discusses the other articles in this Current Issues section and concludes that all of the authors agree that student ratings are valid but that contextual variables such as grading leniency can affect the level of ratings. The authors disagree about the wisdom of applying statistical corrections for such contextual influences. This article argues that the problem lies neither in the ratings nor in the correction but rather in the lack of sophistication of personnel committees who use the ratings. Thus, more attention should be directed toward methods of ensuring more valid use.

Morley (2012): student evaluations in this study were generally unreliable

Morley, D. D. (2012). Claims about the reliability of student evaluations of instruction: The ecological fallacy rides again. *Studies in Educational Evaluation, 38*(1), 15-20. http://dx.doi.org/10.1016/j.stueduc.2012.01.001

[Abstract, abridged] The vast majority of the research on student evaluation of instruction has assessed the reliability of groups of courses and yielded either a single reliability coefficient for the entire group, or grouped reliability coefficients for each student evaluation of teaching (SET) item. This manuscript argues that these practices constitute a form of ecological

correlation and therefore yield incorrect estimates of reliability. Intraclass reliability and agreement coefficients were proposed as appropriate for making statements about the reliability of SETs in specific classes. An analysis of 1073 course sections using inter-rater coefficients found that students using this particular instrument were generally unable to reliably evaluate faculty. In contrast, the traditional ecologically flawed multi-class "group" reliability coefficients had generally acceptable reliability.

Nargundkar & Shrikhande (2012): an instrument that was validated 20 years ago is still valid

Nargundkar, S., & Shrikhande, M. (2012). An empirical investigation of student evaluations of instruction: The relative importance of factors. *Decision Sciences Journal of Innovative Education, 10*(1), 117-135. http://dx.doi.org/10.1111/j.1540-4609.2011.00328.x

[Abstract, abridged] We analyzed over 100,000 student evaluations of instruction over 4 years in the college of business at a major public university. We found that the original instrument that was validated about 20 years ago is still valid, with factor analysis showing that the six underlying dimensions used in the instrument remained relatively intact. Also, we found that the relative importance of those six factors in the overall assessment of instruction changed over the past two decades, reflecting changes in the expectations of the current millennial generation of students. The results were consistent across four subgroups studied—Undergraduate Core, Undergraduate Noncore, Graduate Core, and Graduate Noncore classes, with minor differences.

Rantanen (2013): reliability of SET is questionable; multiple feedbacks required

Rantanen, P. (2013). The number of feedbacks needed for reliable evaluation. A multilevel analysis of the reliability, stability and generalizability of students' evaluation of teaching. *Assessment & Evaluation in Higher Education, 38*(2), 224-239. http://dx.doi.org/10.1080/02602938.2011.625471

[Abstract, abridged] A multilevel analysis approach was used to analyse students' evaluation of teaching (SET). The low value of inter-rater reliability stresses that any solid conclusions on teaching cannot be made on the basis of single feedbacks. To assess a teacher's general teaching effectiveness, one needs to evaluate four randomly chosen course implementations. Two implementations are needed when one course is evaluated, and if one implementation is evaluated, up to 15 feedbacks are needed. The stability of students' ratings is very high, which reflects students' stable rating criteria. There is an obvious rating paradox: from the student's point of view, each rating is very precise, stable and justifiable, but from the teacher's point of view a single feedback reflects the quality of teaching to just a moderate extent. Cross-hierarchical analysis reveals that there are large discrepancies between the uses of rating scales; some students are systematically more lenient in their rating whereas others are systematically more severe. The study also reveals that some courses are generally rated more favourably and that some courses are more suitable for certain teachers.

Socha (2013): a SET instrument was found to have overall good reliability and validity with relatively few biases

Socha, A. (2013). A hierarchical approach to students' assessment of instruction. *Assessment & Evaluation in Higher Education, 38*(1), 94-113. http://dx.doi.org/10.1080/02602938.2011.604713

[Abstract, abridged] Since students are extensively exposed to course elements, students' evaluation of instruction should be one of several components in the teacher evaluation system. Since traditional methods, such as Cronbach's alpha and ordinary least squares regression, do not address the hierarchical data of the classroom, the current study used the statistical techniques of confirmatory factor analysis and hierarchical linear modelling in order to properly investigate the reliability and validity of the Students' Assessment of Instruction (SAI) instrument. Overall, the SAI was found to have good reliability and validity with relatively few biases and could be used to extract five distinguishable traits of instructional effectiveness.

Spooren, Brockx, & Mortelmans (2013): the utility and validity of SET is questionable

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. http://dx.doi.org/10.3102/0034654313496870

[Abstract] This article provides an extensive overview of the recent literature on student evaluation of teaching (SET) in higher education. The review is based on the SET meta-validation model, drawing upon research reports published in peer-reviewed journals since 2000. Through the lens of validity, we consider both the more traditional research themes in the field of SET (i.e., the dimensionality debate, the 'bias' question, and questionnaire design) and some recent trends in SET research, such as online SET and bias investigations into additional teacher personal characteristics. The review provides a clear idea of the state of the art with regard to research on SET, thus allowing researchers to formulate suggestions for future research. It is argued that SET remains a current yet delicate topic in higher education, as well as in education research. Many stakeholders are not convinced of the usefulness and validity of SET for both formative and summative purposes. Research on SET has thus far failed to provide clear answers to several critical questions concerning the validity of SET.

Uttl, White, & Gonzalez (2016): SETs do not indicate teaching quality, meta-analysis

Uttl, B., White, C. A., Gonzalez, D. W. (2016). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation,* (in press, available online September 19, 2106). http://dx.doi.org/10.1016/j.stueduc.2016.08.007

[Abstract, abridged] We re-analyzed previously published meta-analyses of the multisection studies and found that their findings were an artifact of small sample sized studies and publication bias. Whereas the small sample sized studies showed large and moderate correlation, the large sample sized studies showed no or only minimal correlation between SET ratings and learning. Our up-to-date meta-analysis of all multisection studies revealed no significant correlations between the SET ratings and learning. These findings suggest that

institutions focused on student learning and career success may want to abandon SET ratings as a measure of faculty's teaching effectiveness.

Wright & Jenkins-Guarieri (2012): SETs appear to be valid and free from gender bias

Wright, S. L., & Jenkins-Guarieri, M. A. (2012). Student evaluations of teaching: combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education, 37*(6), 683-699. http://dx.doi.org/10.1080/02602938.2011.563279

[Abstract, abridged] Given that there is not one study summarising all these domains of research, a comprehensive overview of SETs was conducted by combining all prior meta-analyses related to SETs. Eleven meta-analyses were identified, and nine meta-analyses covering 193 studies were included in the analysis, which yielded a small-to-medium overall weighted mean effect size (r = .26) between SETs and the variables studied. Findings suggest that SETs appear to be valid, have practical use that is largely free from gender bias and are most effective when implemented with consultation strategies.

Impact on Teaching Quality

Blair & Valdez Noel (2014): little evidence that student feedback is leading to improved teaching

Blair, E., & Valdez Noel, K. (2014). Improving higher education practice through student evaluation systems: is the student voice being heard? *Assessment & Evaluation in Higher Education, 39*(7), 879-894. http://dx.doi.org/10.1080/02602938.2013.875984

[Abstract, abridged] This paper examines the student evaluations at a university in Trinidad and Tobago in an effort to determine whether the student voice is being heard. The research focused on students' responses to the question, 'How do you think this course could be improved?' Student evaluations were gathered from five purposefully selected courses taught at the university during 2011–2012 and then again one year later, in 2012–2013. This allowed for an analysis of the selected courses. Whilst the literature suggested that student evaluation systems are a valuable aid to lecturer improvement, this research found little evidence that these evaluations actually led to any real significant changes in lecturers' practice.

Campbell & Bozeman (2008): questions the effect student evaluations have on teaching quality

Campbell, J. P., & Bozeman, W. C. (2008). The value of student ratings: Perceptions of students, teachers, and administrators. *Community College Journal of Research and Practice, 32*, 13-24. http://dx.doi.org/10.1080/10668920600864137

[Abstract, abridged] This research responded to the lack of emphasis on more effective use of the data for the purpose of improving teaching effectiveness by questioning the opinions and

practices of students, faculty, and administrators. More importantly, this research questioned the value of student ratings of teaching: Is the effort of doing student evaluations worth the institutional investment or is it simply a routine process which has little or no effect on improving teaching?

Curwood, Tomitsch, Thomson, & Hendry (2015): provide an example of support for academics' learning from SETs

Curwood, J.S., Tomitsch, M., Thomson, K., & Hendry. G.D. (2015). Professional learning in higher education: Understanding how academics interpret student feedback and access resources to improve their teaching. *Australasian Journal of Educational Technology, 31*(5). http://dx.doi.org/10.14742/ajet.2516

[Abstract, abridged] Previous research on professional learning has identified that face-to-face consultation is an effective approach to support academics' learning from student feedback. However, this approach is labour and time intensive, and does not necessarily provide all academics with just-in-time support. In this article, we describe an alternative approach, which involves the creation of *Ask Charlie*, a mobile website that visually represents results from student evaluation of teaching (SET), and provides academics with personalised recommendations for teaching resources. *Ask Charlie* was developed and evaluated by drawing on design-based research methods with the aim to support professional learning within higher education.

Makondo & Ndebele (2014): SETs are beneficial for improving teaching quality

Makondo, L., & Ndebele, C. (2014). University lecturers' views on student-lecturer evaluations. *Anthropologist, 17*(2), 377-386. http://www.krepublishers.com/02-Journals/T-Anth/Anth-17-0-000-14-Web/Anth-17-0-000-14-Contents/Anth-17-0-000-14-Contents.htm

[Abstract, abridged] This paper discusses university lecturers' views on student-lecturer evaluation of teaching and learning process. Specific reference is given to the university lecturers' views on the usefulness of the evaluation exercise, the evaluation process, items in the evaluation questionnaires and evaluation feedback reports at a formerly disadvantaged South African University. A total of 118 (53.8%) lecturers out of a staff establishment of 219 teaching staff volunteered their participation in this study. The findings of the study show that insights from student-lecturer evaluations are an important source of information for university teaching staff and administration to consider in their quest to improve on the quality of university teaching and learning moves that can help improve on throughput rates.

Stein, Spiller, Harris, Deaker, & Kennedy (2013): there are gaps in the way academics engage with student evaluation

Stein, S. J., Spiller, D., Terry, S., Harris, T., Deaker, L., & Kennedy, J. (2013). Tertiary teachers and student evaluations: never the twain shall meet? *Assessment & Evaluation in Higher Education, 38*(7), 892-904. http://dx.doi.org/10.1080/02602938.2013.767876

[Abstract, abridged] While extensive research has been done on student evaluations, there is less research-based evidence about teachers' perceptions of and engagement with student evaluations, the focus of the research reported in this paper. Results highlighted the general acceptance of the notion of student evaluations, recurring ideas about the limitations of evaluations and significant gaps in the way academics engage with student evaluation feedback.

Evaluating Faculty for Tenure and Promotion

Boysen (2015): faculty and administrators can over-interpret small variations

Boysen, G. A. (2015). Uses and misuses of student evaluations of teaching: The interpretation of differences in teaching evaluation means irrespective of statistical information. *Teaching of Psychology, 42*(2), 109-118. http://dx.doi.org/10.1177/0098628315569922

[Abstract] Student evaluations of teaching are among the most accepted and important indicators of college teachers' performance. However, faculty and administrators can overinterpret small variations in mean teaching evaluations. The current research examined the effect of including statistical information on the interpretation of teaching evaluations. Study 1 (N = 121) showed that faculty members interpreted small differences between mean course evaluations even when confidence intervals and statistical tests indicated the absence of meaningful differences. Study 2 (N = 183) showed that differences labeled as nonsignificant still influenced perceptions of teaching qualifications and teaching ability. The results suggest the need for increased emphasis on the use of statistics when presenting and interpreting teaching evaluation data.

Boysen, Raesly, & Casner (2014): ratings are misinterpreted by faculty and administrators

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2014). The (mis)interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education, 39*(6), 641-656. http://dx.doi.org/10.1080.02602938.2013.860950

[Abstract, abridged] The current research consisted of three studies documenting the effect of small mean differences in teaching evaluations on judgements about teachers. Differences in means small enough to be within the margin of error significantly impacted faculty members' assignment of merit-based rewards (Study 1), department heads' evaluation of teaching techniques (Study 2) and faculty members' evaluation of specific teaching skills (Study 3). The results suggest that faculty and administrators do not apply appropriate statistical principles when evaluating teaching evaluations and instead use a general heuristic that higher evaluations are better.

Fraile & Bosch-Morell (2015): present a reliable approach to SET interpretation

Fraile, R., & Bosch-Morell, F. (2015). Considering teaching history and calculating confidence

intervals in student evaluations of teaching quality: An approach based on Bayesian inference. *Higher Education, 70*(1), 55-72. http://dx.doi.org/10.1007/s10734-014-9823-0

[Abstract, abbreviated, edited] Student evaluations of teaching quality are among the most used and analysed sources of such information [for lecturer promotion and tenure decisions]. However, to date little attention has been paid in how to process them in order to be able to estimate their reliability. Within this paper we present an approach that provides estimates of such reliability in terms of confidence intervals. This approach, based on Bayesian inference, also provides a means for improving reliability even for lecturers having a low number of student evaluations. Such improvement is achieved by using past information in every year's evaluations.

Jackson & Jackson (2015): concerns with use of SETs for summative purposes

Jackson, M. J., & Jackson, W. T. (2015). The misuse of student evaluations of teaching: Implications, suggestions and alternatives. *Academy of Educational Leadership Journal, 19*(3), 165-173. http://www.alliedacademies.org/academy-of-educational-leadership-journal/

[Abstract, abridged] A five year longitudinal study of the results from Student Evaluations of Teaching (SETs) was accomplished within the business school of a small southwestern state university. Based upon the findings of the study, the authors argue that prior practices in applying the results of SETs for summative purposes have not been based upon a sound statistical foundation. Results from both instructor samples and populations are compared and indicate that the use of means to measure and compare instructor effectiveness requires assumptions of normality which the data does not meet.

Jones, Gaffney-Rhys, & Jones (2015): presents issues if decision-makers use SET results summatively

Jones, J., Gaffney-Rhys, R., & Jones, E. (2014). Handle with care! An exploration of the potential risks associated with the publication and summative usage of student evaluation of teaching (SET) results. *Journal of Further and Higher Education, 38*(1), 37-56. http://dx.doi.org/10.1080/0309877X.2012.699514

[Abstract, abridged] This article presents a synthesis of previous ideas relating to student evaluation of teaching (SET) results in higher education institutions (HEIs), with particular focus upon possible validity issues and matters that HEI decision-makers should consider prior to interpreting survey results and using them summatively. Furthermore, the research explores relevant legal issues (namely, defamation, breach of the duty to take reasonable care for an employee's welfare, breach of the duty of trust and confidence, breach of the right to privacy and, if the lecturer is forced to resign as a consequence of such infringements, constructive dismissal) that decision-makers, in UK HEIs, should appreciate if survey results are widely published or used to inform employment decisions.

Mitry & Smith (2014): conclusions drawn from evaluations may be invalid and harmful

Mitry, D. J., & Smith, D. E. (2014). Student evaluations of faculty members: A call for

analytical prudence. *Journal on Excellence in College Teaching, 25*(2), 56-67. http://celt.miamioh.edu/ject/issue.php?v=25&n=2

[Abstract, abridged] The authors of this article express concern about the use of parametric techniques to report faculty performance based on categorical Likert survey data gleaned from student responses to teaching evaluations. They argue that these surveys often violate primary statistical requirements for evaluative application. Therefore, the conclusions drawn from such evaluations may be invalid and even harmful to faculty members over time. The authors conclude that it is imprudent for university administrators to support questionable analysis methods simply because they have, on the surface, the appearance of rigor, or because the practice has become commonplace.

Palmer (2012): presents examples of ineffective responses to evaluation results

Palmer, S. (2012). Student evaluation of teaching: keeping in touch with reality. *Quality in Higher Education, 18*(3), 297-311. http://dx.doi.org/10.1080/13538322.2012.730336

[Abstract, abridged] This article used publicly available student evaluation of teaching data to present examples of where institutional responses to evaluation processes appeared to be educationally ineffective and where the pursuit of the 'right' student evaluation results appears to have been mistakenly equated with the aim of improved teaching and learning. If the vast resources devoted to student evaluation of teaching are to be effective, then the data produced by student evaluation systems must lead to real and sustainable improvements in teaching quality and student learning, rather than becoming an end in itself.

Multifaceted Evaluation

Berk (2013): covers several issues, including multifactorial evaluations

Berk, R. A. (2013). Top five flashpoints in the assessment of teaching effectiveness. *Medical Teacher, 35*(1), 15-26. http://dx.doi.org/10.3109/0142159X.2012.732247

[Berk is also the author of the 2013 book "Top 10 Flashpoints in Student Ratings and the Evaluation of Teaching"]

[Abstract, abridged] Five flashpoints are defined, the salient issues and research described, and, finally, specific, concrete recommendations for moving forward are proffered. Those flashpoints are: (1) student ratings vs. multiple sources of evidence; (2) sources of evidence vs. decisions: which come first?' (3) quality of ''home-grown'' rating scales vs. commercially-developed scales; (4) paper-and-pencil vs. online scale administration; and (5) standardized vs. unstandardized online scale administrations. Conclusions: Multiple sources of evidence collected through online administration, when possible, can furnish a solid foundation from which to infer teaching effectiveness and contribute to fair and equitable decisions about faculty contract renewal, merit pay, and promotion and tenure.

Cox, Peeters, Stanford, & Seifert (2013): a peer assessment instrument was piloted; formative peer assessment seems important

Cox, C.D., Peeters, M. J., Stanford, B. L., & Seifert, C. F. (2013). Pilot of peer assessment within experiential teaching and learning. *Currents in Pharmacy Teaching and Learning, 5*(4), 311-320. http://dx.doi.org/10.1016/j.cptl.2013.02.003

[Abstract, abridged] Objectives of this study were as follows: (1) to pilot test an instrument for peer assessment of experiential teaching, (2) to compare peer evaluations from faculty with student evaluations of their preceptor (faculty), and (3) to determine the impact of qualitative, formative peer assessment on faculty's experiential teaching. Faculty at Texas Tech University Health Sciences Center School of Pharmacy implemented a new peer assessment instrument focused on assessing experiential teaching. Eight faculty members participated in this pilot. Conclusion: A peer assessment of experiential teaching was developed and implemented. Aside from evaluation, formative peer assessment seemed important in fostering feedback for faculty in their development.

Hughes II & Pate (2013): present a multisource evaluation method

Hughes II, K. E., & Pate, G. R. (2013). Moving beyond student ratings: A balanced scorecard approach for evaluating teaching performance. Issues in *Accounting Education, 28*(1), 49-75. http://dx.doi.org/10.2308/iace-50302

[Abstract, abridged] This position paper proposes a viable alternative to higher education's current focus on student ratings as the primary metric for summative teaching evaluations (i.e., for personnel decisions). In contrast to the divergent opinions among educational researchers about the validity of student ratings, a strong consensus exists that summative measures derived from the student ratings process represent a necessary rather than a sufficient source for evaluating teaching performance (Cashin 1990; Berk 2005). Accordingly, to more completely describe annual teaching performance, we propose a multisource, multiple-perspective Teaching Balanced Scorecard (TBSC), fashioned from the ''classic'' Balanced Scorecard developed by Kaplan and Norton (1992a). The TBSC can guide academic administrators to expand their conceptual view of teaching performance beyond the boundaries of the classroom, while coherently communicating the department's teaching expectations to the faculty; consistent with this proposition, we provide supporting evidence from a successful TBSC implementation in an academic department.

Iqbal (2013): faculty express concerns with peer reviews

Iqbal, I. (2013). Academics' resistance to summative peer review of teaching: questionable rewards and the importance of student evaluations. *Teaching in Higher Education, 18*(5), 557-569. http://dx.doi.org/10.1080/13562517.2013.764863

[Abstract, abridged] This study draws from 30 semi-structured interviews with tenure-track faculty members in a research-intensive university to examine their lack of engagement in the summative peer review of teaching. Findings indicate that most academics in the study do not think peer review outcomes contribute meaningfully to decisions about career advancement

and believe that, in comparison, student evaluation of teaching scores matter more. The findings suggest that faculty member resistance to summative peer reviews will persist unless academics are confident that the results will be seriously considered in decisions about tenure and promotion.

Lyde, Grieshaber, & Byrns (2016): a multisource method of evaluating is a useful tool

Lyde, A.R., Grieshaber, D.C., Byrns, G. (2016). Faculty teaching performance: Perceptions of a multi-source method for evaluation (MME). *Journal of the Scholarship of Teaching and Learning, 16*(3), 82-94. http://dx.doi.org/10.14434/josotl.v16i3.18145

[Abstract, abridged] A holistic system of evaluating university teaching is necessary for reasons including the limitations of student evaluations and the complexity of assessing teaching performance. University faculty members were interviewed to determine their perceptions of the multisource method of evaluating (MME) teaching performance after a revision of policies and procedures was approved. The MME is comprised of three primary data sources: student evaluations, instructor reflections describing attributes of their own teaching such as the teaching philosophy, and a formative external review. While the faculty perceived the MME as a useful tool, they still believe it operates more to produce a summative product than work as a formative process. According to the results, a more formative process would be supported by addressing several factors, including timing of reflections, accountability from year to year, and mentoring. Improving these constraints may make the proposed MME a more appropriate tool for formative review of teaching.

Marsh & Roche (1997): multidimensional aspects of teaching should be evaluated; suggest nine factors

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187-1197. http://dx.doi.org/10.1037/0003-066X.52.11.1187

This article has been included in previous themes. For this theme, Marsh & Roche (1997) believe that effective evaluation tools should consider nine factors: "Learning/Value, Instructor Enthusiasm, Organization/Clarity, Group Interaction, Individual Rapport, Breadth of Coverage, Examinations/Grading, Assignments/Readings, and Workload/Difficulty" (p.1187). The authors also comment on the nature of "homemade" evaluation instruments being of questionable quality (p. 1188).

Martin, Dennehy, & Morgan (2013): validity of SET is questioned; student focus groups suggested as an alternative

Martin, L. R., Dennehy, R., & Morgan, S. (2013). Unreliability in student evaluation of teaching questionnaires: Focus groups as an alternative approach. *Organization Management Journal, 10*(1), 66-74. http://dx.doi.org/10.1080/15416518.2013.781401

[Abstract, abridged] Research on the validity and reliability of SETs is vast, though riddled

with inconsistencies. The many “myths” of SETs are investigated and the incongruities are demonstrated. We hypothesize that the discrepancies in empirical studies come from misunderstanding and inappropriate actions by students. To address the complexity inherent in these problems, we suggest the use of focus groups as an alternative approach or complement to the standard SETs. A recommended format and guidelines for running classroom focus groups are provided. Institutional constraints and implementation concerns are addressed as well. This article lays the foundation for implementing a change in student assessment of teaching by proposing a method to compensate for bias in SETs, using focus groups as an evaluation tool, either as a stand-alone process or as a supplement to current methods.

Ridley & Collins (2015): suggests a comprehensive performance evaluation instrument

Ridley, D., & Collins, J. (2015). A suggested evaluation metric instrument for faculty members at colleges and universities. *International Journal of Education Research, 10*(1), 97-114. Retrieved from http://eds.a.ebscohost.com.login.ezproxy.library.ualberta.ca/eds/pdfviewer/pdfviewer?sid=9ff24389-d34d-43d1-83fc-6ef82bd1ad47%40sessionmgr4009&vid=2&hid=4102

[Abstract, abridged] This study puts forth a comprehensive performance evaluation method for university faculty members. The instrument is comprised of a teaching evaluation metric, a research evaluation metric, and a service evaluation metric. This study provides a unique method for measuring the performance of university faculty members by regressing cumulative student grade point average on the fraction of the total number of credit hours that students are taught by each faculty member. The study postulates that the resulting regression coefficients measure the average rate at which each faculty member contributes to student learning as measured by cumulative grade points earned per contact hour of instruction. Since this model of teaching effectiveness is based on grades, freely assigned by individual faculty members, it is a no contact, non-intrusive, non-confrontational, non-threatening, non-coercive evaluation of teaching.

Stupans, McGuren, & Babey (2016): present a tool for analyzing free-form comments on ratings forms

Stupans, I., McGuren, T., & Babey, A. M. (2016). Student evaluation of teaching: A study exploring student rating instrument free-form text comments. *Innovative Higher Education, 41*(1), 33-52. http://10.1007/s10755-015-9328-5

[Abstract] Student rating instruments are recognised to be valid indicators of effective instruction, providing a valuable tool to improve teaching. However, free-form text comments obtained from the open-ended question component of such surveys are only infrequently analysed comprehensively. We employed an innovative, systematic approach to the analysis of text-based feedback relating to student perceptions of and experiences with a recently developed university program. The automated nature of the semantic analysis tool "Leximancer" enabled a critical interrogation across units of study, mining the cumulative text for common themes and recurring core concepts. The results of this analysis facilitated the identification of issues that were not apparent from the purely quantitative data, thus providing

a deeper understanding of the curriculum and teaching effectiveness that was constructive and detailed.

[Link from **Zimmerman** (2008): some tools may encourage students to focus on negative aspects of teaching; anonymous feedback means that students are not held accountable for their comments

Zimmerman, B. (2008). Course evaluations - students' revenge? *University Affairs.* Retrieved from http://www.universityaffairs.ca/opinion/in-my-opinion/course-evaluations-students-revenge/

This is an online opinion article.

"Even choosing the right questions is difficult. Instead of 'What did you like least about the lectures?' shouldn't we be asking, 'Is there something you liked least about the lectures?' When we manipulate students into providing negative responses, we encourage them to cast about for some negative remark, *any* negative remark, when they might otherwise have been declined" (paragraph 7).

"Many students don't need any encouragement to bash their teachers. The exercise is meant in part to ensure that instructors are held accountable, yet students engage in libel with impunity. The student who referred to a colleague as a "cow" was not held accountable" (paragraph 8).

Appendix I: Recommendations Related to Evaluation of Teaching from the 2013 Renaissance Committee Report

These recommendations are taken from pages 11 and 12 of the report.

Source: Cheeseman, C., MacLaren, I., Carey, J., Glanfield, F., Liu, L., McFarlane, L., Cahill, J. C., Garneau, T., Supernant, K., & Szeman, I. (2013, December 9). *Report of the Renaissance Committee.* Retrieved from http://www.renaissance.ualberta.ca/

3-2 That all scholars be evaluated using the same evaluation structure, with constituency-specific evaluation committees. Non-scholarly activities should be evaluated separately.

3-3 That the number of committees evaluating the excellence of scholarly activities performed by a single constituency be substantially reduced from 3 to 6. Such committees will be formed around scholarly discipline, not faculty boundaries. Cultural practices within the unit should not be allowed to influence the salary trajectories nor the process by which scholars are evaluated.

3-4 That there be greater consistency in the size of comparator groups used for evaluation, at both the small and large unit levels.

3-8 That all scholars, which include tenure-track faculty, librarians, and specialized scholars, be evaluated in accordance with the broad definition of Scholarship provided in Section 2 of this report. These constituencies should be evaluated equitably based on the Scholarship performance measures and the extent to which Scholarship comprises a part of their duties.

3-9 That all scholarly activities be evaluated using more than simple metrics (e.g. Impact Factors, USRI); that multifaceted evaluations be applied to all scholarly activities to allow for identification of scholarly excellence.

3-11 Establishment of a Teaching Strategy for the University of Alberta that reviews and updates the teaching and learning policies currently in place in the GFC Policy Manual, and determined implementation of those policies.

3-12 Creation of specific, transparent policies for teaching evaluation to guide annual reviews, contract renewal decisions, and decisions on tenure and promotion. (As, for example, delineated in the CAUT model policy on the evaluation of teaching performance, create policies and procedures that allow recognition of all aspects of teaching duties performed by academic staff.)

3-13 Establish a committee to redesign the USRI questions, ensuring a reliable and valid tool that meets international standards for summative evaluation, provides a degree of formative feedback, minimizes the potential for derogatory feedback, ensures value to the students who

participate in the process, and is in alignment with the University's Teaching Strategy. To ensure movement on this recommendation, establish a two-year limit on implementation.

3-14 If changes to the USRI are not accomplished within two years (end of Fall term, 2015), (AASUA and Administration) declare a moratorium on their use.

3-15 Provide leadership, support, and resources further to encourage teaching development and teaching Scholarship at the University of Alberta.

3-16 Standardize reporting periods for all evaluation committees.

3-22 require all scholarly evaluation committees to use external standards for the assessment of Scholarship, reaching decisions by reference to agreed-upon external standards rather than to colleagues' performance.