Student Ratings of Instruction: A Summary of the Literature

By: Brianna Strumm

Prepared for: The USRI Working Group at the University of Calgary

May 17, 2019

Summary on the Research of Student Evaluations

This report on student evaluations was conducted throughout January - April 2019 to support the work of the Universal Student Ratings of Instruction (USRI) working group looking at the renewal of the USRI questionnaire at the University of Calgary. This report was informed by approximately 115 articles from scholarly journals. Of this work, approximately 75 articles were integrated here. Items were included based on the relevance and suitability of the research to the project - specifically the year it was published and where it was conducted. The majority of the research articles surveyed were published over a 20-year span, between 1999 and 2019 primarily in North America. This summary provides a comprehensive (not exhaustive) look at the literature and is intended to provide a balanced view within a current Canadian narrative on this issue.

This review has also been informed by reports from other institutions that have done previous work on this topic (such as the University of Toronto and the University of Saskatchewan), as well as the arbiration decision between Ryerson University vs. Ryerson Faculty Association (2018), and the Report of the OCUFA Student Questionnaires on Courses and Teaching Working Group (2019). These reports provided helpful insights on current issues surrounding the information collected and the use of student ratings of instruction as well as supplemental overviews of the literature. A few editorial and opinion pieces relevant to current issues and trends institutions are facing regarding student ratings were also included. These included recent pieces from University Affairs and The Chronicle of Higher Education.

This summary is broken into four sections. The first section describes the significant insights derived from the literature, including current challenges associated with student ratings of instruction. Section two highlights different issues associated with student ratings processes found most commonly in the literature. The third section describes the few recommendations from the literature on how to improve this process. Section four addresses the system of teaching evaluation and describes the current conversations happening in this domain. For the sake of brevity, it is assumed that the reader of this report understands the meaning of universal student ratings of instruction (USRI or SRI) that are conducted at most post-secondary institutions, including the University of Calgary.

1. Research Insights and Challenges

In North America, the practice of obtaining student feedback on individual instructors is widespread (Richardson, 2005). According to the literature, a fundamental purpose of collecting student feedback is to attain diagnostic, formative feedback to teachers when looking back over their teaching (Marsh, 2007; Marsh & Dunkin, 1992; Marzano, 2012). After nearly three decades of research in this field, Marsh (2007) has concluded that when appropriately designed and used, student feedback questionnaires can be a helpful measurement tool for formative purposes (Marsh, 2007).

There is a lack of Canadian data on the topic of SRIs. Most studies on student ratings are conducted in the United States. While there are some similarities between the higher education in Canada and the United States, there are also significant differences in terms of structure, organization and accountability measures, not to mention cultural and demographic variations (Gravestock & Gregor-Greenleaf, 2008).

Furthermore, institutional policies and practices vary between these two countries. Therefore, there are limitations in which these findings can be generalized across the Canadian context.

This summary of the literature also revealed disagreement among scholars. As pointed out by Gravestock and Gregor-Greenleaf in 2008 at the University of Toronto, even when issues such as reliability, validity and utility have been resolved and supported by strong research, subsequent investigative studies arise which further complicates matters. Furthermore, authors highlight and reference studies that have been refuted or are ungeneralizable due to methodological challenges such as size and scope of a study. Other scholars, such as Benton and Ryalls (2016), speak more broadly to the misconceptions and myths that exist around student ratings of instruction, in order to refute some of the claims about bias, response rates, correlation to student grades, and so on, that have been made by smaller-scale studies.

As such, drawing generalizable conclusions from research studies is challenging, at best. The size and scope of studies of student ratings of instruction are relatively small and context-specific; the literature consists of an amalgamation of single-course or single-institution case studies or mixed methods studies. This makes it difficult to read across the data to extrapolate any generalizable or transferrable findings. Limited meta-analysis is present within the literature, especially from a Canadian perspective. Another challenge to applying research findings is the diversity of student ratings instruments, policies and processes, as well as the diversity of institutional and instructional contexts. These all vary significantly across and within institutions. Considering the context-dependent nature of examining student evaluations at institutional levels, it's important to remain cautious about making any conclusions when presented with this overview of literature.

2. Factors to Consider When Designing and Administering Student Ratings Instruments

Nira Hativa (2013) has spent years summarizing the literature on the topic of SRIs. She suggests that student ratings can serve both an important practical and theoretical purpose by providing educators and researchers with information on the teaching-learning process in a particular context. Information gathered may be helpful when assessing the usefulness of new, innovative pedagogical techniques or in judging the impact of instructional strategies for different students, courses, and settings (Abrami, d'Apollonia & Rosenfield, 2007). Formative evaluation can also be beneficial when revealing the exact nature of teaching strengths and difficulties. In order to learn from their experiences, instructors require specific feedback about where they have been successful and where they may have fallen short (Andrade & Cizek, 2010; Hattie & Gan, 2011). Our summary of the literature revealed some of these highly debated non-instructional factors when designing, administering and reporting student ratings instruments. Institutions cannot necessarily control for these factors at an administrative level, however, these factors can influence outcomes of students' ratings and need to be taken into consideration when building a student questionnaire.

2.1 Gender and instructor characteristics

The influence of gender on student ratings of instruction is multifaceted. There is, however, strong suggestion in the literature that the gender and characteristics of instructors influence the ways students evaluate instruction. Evidence of gender bias and other forms of bias related to instructor characteristics

has been found in multiple individual studies (Arbuckle, J. & Williams, 2003; Boring, Ottoboni & Stark, 2017; Boring, 2017; Centra & Gaubatz, 2000; MacNell, Driscoll, & Hunt, 2015; Mengel, Sauermann & Zolitz, 2017; Miller & Chamberlin, 2000; Sprague & Massoni, 2005). From this research, certain types of questions clearly elicit and lead to bias, such as questions about instructor knowledge or questions that ask students about specific instructor characteristics or behaviors (Boring, 2017; Miller & Chamberlain, 2000). In some cases, "these biases may pull ratings in multiple directions for any given pairing of instructor and student" (Greenleaf & Gravestock, 2017, p. 3). Despite some statistical limitations in the literature (Benton & Cashin, 2014; Benton & Ryalls, 2016; Hativa, 2013; DeFrain, 2016; Spooren, Brockx, & Mortelmans, 2013), there is sufficient evidence to point to gender biases and gendered expectations of their instructors.

Sprague and Massoni (2005) and Laube, Massoni, Sprague and Ferber (2007) found that students expected women faculty to be caring and nurturing and expected male faculty to be entertaining and personable. Women faculty are typically praised for being empathetic, approachable and for fostering a good relational climate in the classroom (Bachen, McLoughlin & Garcia, 1999; Basow & Montgomery, 2005; Basow, Phelan & Capotosto, 2006; Centra & Gaubatz, 2000). Men, on the other hand receive higher ratings on dimensions such as course planning, competence, knowledge, and organization skills (Basow et al., 2006). Research has shown that female instructors who do not conform to 'traditional' feminine gender roles tend to be perceived negatively by both male and female students (Bachen et al., 1999; Basow & Montgomery, 2005; Basow et al., 2006); Sprague and Massoni (2005) also found that female faculty who do not meet gendered expectations are more at risk of getting less favorable ratings from students. These studies are based on both qualitative data analysis of student comments as well as larger-scale questionnaires asking students about specific characteristics displayed by their instructors.

As mentioned above, questions that ask students to rate particular instructor characteristics or behaviors is shown to be responsible for the some of the evidence of instructor bias. There is some evidence to suggest that biases around gender, ethnicity, language and even personality characteristics such as "enthusiasm" or "charisma" factor into students' ratings of their instructors. These biases mean that students' ratings can be, in some cases, discriminatory against vulnerable faculty members who do not display certain traits such as 'high energy' or 'positive self-esteem' (Benton & Cashin, 2012). Studies on gender bias are commonly undertaken from perspectives that are more sensitive to intersectional identities – thus confounding the picture of which characteristic or trait is the one that activates students' biases. Very few studies address single instructor characteristics in isolation (Lazos, 2012; Pittman, 2010; Ryalls & Benton, 2017).

2.2 Learning Environment

Learning environment is another well-researched topic in studying student evaluation of teaching. Some studies comparing student ratings in distance education and on-campus courses concluded that students in both instructional modes gave similar ratings of overall course and instructional quality (Carle, 2009; Kelly, Ponton & Rovai, 2007; McGhee & Lowell 2003; Waschull, 2001). One study, however, stated that instructors and administrators in higher education should recognize that instructors may receive fewer positive scores for their online courses (Young & Duncan, 2014). This may be explained by difficulties in the following areas: communication, creating positive faculty-student interactions, establishing agreeable

and fair grading criteria, influencing student beliefs about their own learning, encouraging student effort and finding useful online teaching strategies for student satisfaction (Young & Duncan, 2014).

2.3 Class size

The relationship between class size and student ratings is also conflicted. Some studies found a relationship between favorable student ratings and smaller classes (Algozzine et al., 2004; Badri, Abdulla, Kamali & Dodeen, 2006; Feldman, 2007; Isley & Singh, 2007; Koh & Tan, 1997; Liaw & Goh, 2003). Although these aforementioned studies have found smaller classes often receive slightly higher ratings, the correlation between class size and ratings is statistically insignificant and is therefore not viewed as having any impact on validity of the instruments used (Cashin, 1995; ; d'Apollonia & Abrami, 1997; Marsh & Roche, 1997; McKeachie, 1997). The proposed explanation by these authors regarding more favorable ratings of small classes is that when the class size is small, students tend to have increased opportunities for student-instructor interaction and critical-thinking activities. Because instructors may not have much agency over class size, care should be taken to contextualize class size in data reports (Gravestock & Gregor-Greenleaf, 2008).

2.4 The halo effect

The halo effects describes the way that initial impressions of an instructor can have a lasting effect on students' perceptions of teaching. Interest in determining at what point students form lasting impressions of the instructor is not a new question. Previous results by Buchert, Laws, Apperson and Bregman (2008) suggest that students form an opinion about the instructor during the first two weeks of the semester and that these opinions are reflected in ratings at the end of the semester (the halo effect). Students' first impressions of their instructors the first day of class appears to significantly influence students' end-of-semester ratings of instruction (Buchert et al. 2008). Findings suggest that if students have a particularly powerful positive or negative opinion of one of the teacher's characteristics, then they are likely to give an overall rating that is more positive or negative based on that one opinion (Keeley, English, Irons & Henslee, 2013). Previous researchers have also examined the effect of information received before being exposed to the instructor on subsequent ratings of the instructor - otherwise knowns as instructor reputation. One study concluded that when the instructor's reputation was manipulated in the form of positive or negative primes (warm or cold), it affected subsequent student ratings of that instructor (Griffin, 2001; Wheeler, Wright & Frost, 2005).

2.5 Course type, level and discipline

While this topic was not comprehensively examined within the literature, some studies point to students frequently ranking electives (classes they choose) somewhat more positively than required courses (Algozzine et al., 2004; Marsh & Roche 1997). However, this has not been found to have a statistically significant impact on ratings. Graduate students tend to rate instructors more favorably than undergraduate students (Marsh, 2007; Whitworth, Price & Randall, 2002) as graduate students are generally expected to have a greater liking for the subject than undergraduates (Nargundkar & Shrikhande, 2014). A possible explanation is that more inexperienced undergraduate students may not have an accurate understanding of what university-level courses should be like, and therefore they are less

familiar with the learning expectations than students in upper level courses are (Nargundkar & Shrikhande, 2014). Moreover, some studies have shown that discipline can play a part in how instructors are being rated. The arts and humanities tend to receive the highest ratings, followed by the social sciences, and then natural sciences (Centra, 2009; DeFrain, 2016; Neumann, 2001; Ory, 2001). Although there is some evidence that ratings differ between disciplines, it is not clear why. This may reflect disciplinary difference in teaching styles and goals rather than bias; others attribute it to student background preparation or subject-matter difficulty (Benton & Ryalls, 2016). However, more recent studies suggest that the overall nature of effectual teaching and learning are nevertheless quite uniform across disciplines (Hativa, 2013).

3. Recommendations for Improving Student Questionnaires and Processes

The literature makes recommendations on ways to improve this process. Recommendations are made in the following areas including: a focus on student learning, streamlining administration and ensuring a more comprehensive undertaking of looking at teaching.

3.1 Focus on the Learning Experience

Some researchers have argued that evaluations are best suited to measure a student's experience of their own learning and suggest that the most appropriate criterion for looking at teaching is whether students learn and how they best learn (Marsh & Roche, 2000; Murray, 2005; Stark & Frishtat, 2014; Svinivki & McKeachie, 2010). Historically, student ratings tend to be limited in their ability to do that (Charbonneau, 2013). Student ratings typically assess the student's subjective experience/satisfaction with an instructor's behavior and/or their characteristics, and not how an instructor has shaped student learning with the context of a course. In a comprehensive meta-analysis of highly-cited publications, Uttl, White and Gonzalez (2017) found that contemporary student rating instruments are more likely to measure student satisfaction with a course rather than their learning experience.

The University of Oregon, for example, has students select, from a list, teaching elements that were most beneficial to their learning and those that could use some improvement. They were then asked to provide written comments about those areas. The responses are aggregated, so professors can see if a cluster of comments indicating particular weaknesses or strengths within the design of the course that led to optimal learning. The goal of all of those efforts is not only to minimize punitive non-instructional variables but also to ensure that instructors can learn from student feedback and respond accordingly (Doerer, 2019).

3.2 Online Administration

Benton and Cashin (2012) summarized the conclusions of the major reviews of the student ratings research and literature from the 1970s to 2010. They found in their review that the quality of student feedback when administered through an online platform appears to be similar to paper-based administration with no significant differences in scores (Benton & Cashin, 2012). Results from a pilot study of 10,417 students registered in 318 courses at the University of Ottawa suggest an average decrease in participation rate of 12-15% when using an online system (Groen & Herry, 2017). Instructors

and students alike suggested that an in-class period be maintained for the electronic completion of student ratings (Groen & Herry, 2017).

One of the positive aspects of online administration of student feedback is the quality and quantity of comments from students regarding their educational experience (Crews & Curtis, 2011; Hativa, 2013; Legg & Wilson, 2012; Venette, Sellnow & McIntire, 2010). Importantly, the length of written comments is greater when students are typing online than writing in the classroom. Part of the difference may be due to the fear of having their handwriting recognized by their instructor on in-class forms (Avery, Bryant, Mathios, Kang & Bell, 2006). Or, it may be that students have more time to think about their responses when typing them online. Despite the potential of a lower response rate, online instruments yielded five times the total amount of written commentary (Hardy, 2003). The research overall shows promising results in support of online administration, which in turn also offer numerous advantages, notably ease of administration and completion, speed of analysis and reporting and complete confidentiality of comments.

3.3 Assessing Teaching

Given the limitations, biases, and complexities referenced in this report, it is clear that student ratings are not always a direct measure of the actual teaching. With this in mind, other measures are needed to assess teaching, such as self-assessments or peer feedback (Weschke & Canipe, 2010). In addition to end-of-course student ratings, other formative and summative measures suggested within the literature include: structured peer classroom observation, peer review of course materials, video classroom review, teaching scholarship, teaching portfolio review and student outcomes (Baldwin & Blattner, 2003; Berk, 2018; Sproule, 2002).

Including additional metrics and processes as suggested above are “designed to provide a 360-degree view so that a comprehensive picture of the faculty member's accomplishments and areas for improvement could be obtained” (Weschke & Canipe, 2010, p. 46). Student feedback and ratings could also be used for other activities such as improving graduate or professional student supervision. Some instructors might be less well perceived in class, but perform well in one-to-one interactions (Arah, Heineman & Lombarts, 2012). Using a variety of alternative methods to provide insight into teaching may also minimize the effect of bias on outcomes for instructors. Furthermore, some institutions implement a midterm student-experience survey that only the applicable instructor can view. An instructor can then make changes in the middle of a semester, when students can still benefit from improvements, encouraging them to give constructive and formative feedback (Doerer, 2019).

Due to the limited access to Canadian institutional data on this topic and the complexity of existing knowledge and practice, the USRI working group recommends further research on analyzing teaching and learning trends locally at the University of Calgary. There is certainly an opportunity to explore this topic specifically from a Canadian institutional standpoint, should this be of interest to the University of Calgary. There is limited research which captures the experiences with gender and racial bias in Canadian institutions, as well as the impacts of course ratings with Indigenous scholars. Furthermore, this summary has propelled the working group into wanting to make sense of the localized context and appropriately move forward with revamping the USRI questionnaire at the University of Calgary. As such, any claim of knowledge generated from this summary is equally matched with a need for further inquiry and curiosity

as we collaborate with the University of Calgary community on building the values and principles for attaining course feedback.

References

Abrami, P., d'Apollonia, S., & Rosenfield, S. (2007). The dimensionality of student ratings of instruction: what we know and what we do not. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence based perspective* (pp. 385–456). Dordrecht, The Netherlands: Springer.

Algozzine, B., Gretes, J. Flowers, C. Howley, L. Beattie, J., Spooner, F., … Bray, M. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching, 52*(4), 134-141. https://doi.org/10.3200/CTCH.52.4.134-141

Andrade, H. L., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment.* New York: Routledge.

Apodaca, P. & Grad, H. (2005). The dimensionality of student ratings of teaching: Integration of uni- and multidimensional modes. *Studies in Higher Education, 30*(6), 723-748. https://doi.org/10.1080/03075070500340101

Arah, A. O., Heineman, M. J., & Lombarts, Kiki M. J. M. H. (2012). Factors influencing residents' evaluations of clinical faculty member teaching qualities and role model status. *Medical Education, 46*, 381-389. https://doi.org/10.1111/j.1365-2923.2011.04176.x

Arbuckle, J. & Williams, B. (2003). Students' perceptions of expressiveness: Age and gender effects on teaching evaluations. *Sex Roles, 49*(9-10), 507-516. https://doi.org/10.1023/A:1025832707002

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations: Does an online delivery system influence student evaluations? *Journal of Economic Education, 37,* 21-37. https://doi.org/10.3200/JECE.37.1.21-37

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education, 48*, 193–210. https://doi.org/10.1080/03634529909379169

Badri, M., Abdulla, M., Kamali, M., & Dodeen, H. (2006). Identifying potential biasing variables in student evaluation of teaching in a newly accredited business program in the UAE. *International Journal of Educational Management, 20*, 43–59. https://doi.org/10.1108/09513540610639585

Baldwin, T. & Blattner, N. (2003). Guarding against potential biases in student evaluations: What every faculty member needs to know. *College Teaching, 51*(1), 27-32. https://doi.org/10.1080/87567550309596407

Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-rating of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education, 18*, 91–106. https://psycnet.apa.org/doi/10.1007/s11092-006-9001-8

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly, 30*, 25–35. https://doi.org/10.1111/j.1471-6402.2006.00259.x

Benton S. L., Cashin, W. E. (2012). *Student ratings of teaching: A summary of research and literature. (IDEA Paper No. 50).* Manhattan, KS: The IDEA Center. Retrieved from https://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA_50.pdf

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In M. B. Paulsen (Ed.), *Higher education: Handbook of theory & research*, (Vol. 29, pp. 279-326). Dordrecht, The Netherlands: Springer.

Benton, S. L. & Ryalls, K. R. (2016). *Challenging misconceptions about student ratings of instruction*

(IDEA Paper No. 58). Manhattan, KS: The IDEA Center. Retrieved from https://www.ideaedu.org/Portals/0/Uploads/Documents/IDEA%20Papers/IDEA%20Papers/PaperIDEA_58.pdf

Berk, R. (2018). Start spreading the news: Use multiple sources of evidence to evaluate teaching. *The Journal of Faculty Development, 32*(1), 73-81.

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145*, 27-41. https://doi.org/10.1016/j.jpubeco.2016.11.006

Boring, A., Ottoboni, K., & Stark, P. B. (2017). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *Science Open Research, 2016*(01), 1-11.

Buchert, S., Laws, E. L., Apperson, J. M., & Bregman, N. J. (2008). First impressions and professor reputation: Influence on student evaluations of instruction. *Social Psychology of Education, 11*(4), 397–408. https://psycnet.apa.org/doi/10.1007/s11218-008-9055-1

Carle, A. C. (2009). Evaluating college students' evaluations of a professor's teaching effectiveness across time and instruction mode (online vs. face-to-face) using a multilevel growth modeling approach. *Computers & Education, 53*(2), 429–435. https://doi.org/10.1016/j.compedu.2009.03.001

Cashin, W. E. (1995). *Student ratings of teaching: The research revisited. (IDEA Paper No. 32.)* Manhattan, KS: The IDEA Center. Retrieved from https://files.eric.ed.gov/fulltext/ED402338.pdf

Centra, J. A. (2009). *Differences in responses to the Student Instructional Report: Is it bias?* Princeton, NJ: Educational Testing Service.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education, 71*, 17–33. https://doi-org.ezproxy.lib.ucalgary.ca/10.1080/00221546.2000.11780814

Charbonneau, L. (2013, August). Course evaluations: the good, the bad and the ugly. *University Affairs*, Retrieved from https://www.universityaffairs.ca/features/feature-article/course-evaluations-the-good-the-bad-and-the-ugly/

Crews, T. & Curtis, D. (2011). Online course evaluations: Faculty perspective and strategies for improved response rates. *Assessment & Evaluation in Higher Education, 36*(7), 865-878. https://doi.org/10.1080/02602938.2010.493970

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

DeFrain, E. (2016). *An analysis of differences in non-instructional factors affecting teacher-course evaluations over time and across disciplines.* (Doctoral dissertation, University of Arizona, Arizona).

Doerer, K. (2019, January). Colleges are getting smarter about student evaluations. Here's how. *The Chronicle of Higher Education.* Retrieved from https://www.chronicle.com/article/Colleges-Are-Getting-Smarter/245457

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93-129). Dordrecht, The Netherlands: Springer.

Gravestock, P. & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends.* Toronto, ON: Higher Education Quality Council of Ontario. Retrieved from http://www.heqco.ca/SiteCollectionDocuments/Student%20Course%20Evaluations_Research,%20Models%20and%20Trends.pdf

Greenleaf, E. & Gravestock, P. (2017). *Further research on gender bias in course evaluation data.* Retrieved from http://www.artsci.utoronto.ca/faculty-staff/teacher-info/atlas/Genderbiasincourseevaluations.pdf

Griffin, B. W. (2001). Instructor reputation and student ratings of instruction. *Contemporary Educational Psychology, 26,* 534-552. https://doi.org/10.1006/ceps.2000.1075

Groen, J. F. & Herry, Y. (2017). The online evaluation of courses: Impact on participation rates and evaluation scores. *Canadian Journal of Higher Education, 47*(2), 106-120. Retrieved from https://files.eric.ed.gov/fulltext/EJ1154163.pdf

Hardy, N. (2003). Online ratings: Fact and fiction. *New Directions for Teaching and Learning, 96,* 31-38. https://doi.org/10.1002/tl.120

Hattie, J., & Gan, M. (2011). Instruction based on feedback. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 249–271). New York: Routledge.

Hativa, N. (2013). *Student ratings of instruction: Recognizing effective teaching.* Charleston, SC: Oron Publications.

Isley, P., & Singh, H. (2005). Do higher grades lead to favorable student evaluations? *Journal of Economic Education, 36*, 29– 42. https://doi.org/10.3200/JECE.36.1.29-42

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. https://doi.org/10.1177%2F0013164412475300

Kelly, H. F., Ponton, M. K., & Rovai, A. P. (2007). A comparison of student evaluations of teaching between online and face-to-face courses. *Internet and Higher Education, 10*, 89–101. https://doi.org/10.1016/j.iheduc.2007.02.001

Koh, H. C., & Tan, T. M. (1997). Empirical investigation of the factors affecting SET results. *International Journal of Educational Management, 11*(4), 170–178. https://doi.org/10.1108/09513549710186272

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*, 87–104. Retrieved from http://www.jstor.org/stable/40071230

Lazos, S. R. (2012). Are student teaching evaluations holding back women and minorities? The perils of "doing" gender and race in the classroom. In G. G. y Muhs, Y. F. Niemann, C. G. Gonzalez, & A. P. Harris (Eds.), *Presumed Incompetent: The intersections of race and class for women in academia* (pp. 164–185). Boulder, CO: University Press of Colorado.

Legg, A. M., & Wilson, J. H. (2012). RateMyProfessors.com offers biased evaluations. *Assessment & Evaluation in Higher Education, 37*(1), 89–97. https://doi.org/10.1080/02602938.2010.507299

Liaw, S., & Goh, K. (2003). Evidence and control of biases in student evaluations of teaching. *The International Journal of Educational Management, 17*(1), 37–43. https://doi.org/10.1108/09513540310456383

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

Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence based perspective* (pp. 319–384). New York: Springer.

Marsh, H. W. & Dunkin, M. J. (1992). Students' evaluations of university teaching: a multidimensional Perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research*, Volume 8 pp. New York: Agathon Press.

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. https://psycnet.apa.org/doi/10.1037/0003-066X.52.11.1187

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: Popular myth, bias, validity, and innocent bystanders. *Journal of Educational Psychology, 92*, 202-22. https://psycnet.apa.org/doi/10.1037/0022-0663.92.1.202

Marzano, R. J. (2012). The Two Purposes of Teacher Evaluation. *Educational Leadership, 70*(3), 14-19.

McGhee, D. E., & Lowell, N. (2003). Psychometric properties of student ratings of instruction in online and on-campus courses. *New Directions for Teaching & Learning, 96*, 39–48. https://doi.org/10.1002/tl.121

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*(11), 1218-1225.

Mengel, F., Sauermann, J. & Zolitz, U. (2017). *Gender bias in teaching evaluations.* Maastrict, The Netherlands: Research Centre for Education and the Labour Market.

Miller, J. & Chamberlin, M. (2000). Women are teachers, men are professors: A study of student perceptions. *Teaching Sociology, 28*, 283-298. http://dx.doi.org/10.2307/1318580

Murray, H. G. (2005, June). *Student evaluation of teaching: Has it made a difference?* Paper presented at the Annual Meeting of the Society for Teaching and Learning in Higher Education, Charlottetown, Prince Edward Island, Canada.

Nargundkar, S. & Shrinkhande, 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. https://doi.org/10.1111/j.1540-4609.2011.00328.x

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

Neumann, R. (2001). Disciplinary differences and university teaching. *Studies in Higher Education, 26*(2), 135-146. https://doi.org/10.1080/03075070120052071

Ontario Confederation of University Faculty Associations Working group. (2019), *Report of the OCUFA Student Questionnaires on Courses and Teaching*, <https://ocufa.on.ca/assets/OCUFA-SQCT-Report.pdf>, retrieved on May 20, 2019.

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. *New Directions for Teaching and Learning, 87,* 3-15. https://doi.org/10.1002/tl.23

Pittman, C. T. (2010). Race and Gender Oppression in the Classroom: The Experiences of Women Faculty of Color with White Male Students. *Teaching Sociology, 38*(3), 183–196. https://doi.org/10.1177%2F0092055X10370120

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature.

Assessment & Evaluation in Higher Education, 30(4), 387-415. https://doi.org/10.1080/02602930500099193

Ryalls, K. & Benton, S. (2017). Myths and misconceptions of student ratings: Gender bias and more. Retrieved from https://www.ideaedu.org/Resources-Events/IDEA-Blog/PostId/46/myths-and-misconceptions-of-student-ratings-gender-bias-and-more

Ryerson University v Ryerson Faculty Association, 2018. CanLII 58446 (ON LA), <http://canlii.ca/t/hsqkz>, retrieved on 2019-05-22

Sprague, J., & Massoni, K. (2005). Student evaluations and gendered expectations: What we can't count can hurt us. *Sex Roles, 53,* 779–793. https://doi.org/10.1007/s11199-005-8292-4

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. https://doi.org/10.3102%2F0034654313496870

Sproule, R. (2002). The underdetermination of instructor performance by data for the student evaluation of teaching. *Economics of Education Review, 21*(3), 287-297. https://doi.org/10.1016/S0272-7757(01)00025-5

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

Svinivki, M., & McKeachie, W. J. (2010). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers*. Boston, MA: Houghton Mifflin.

Venette S., Sellnow D. & McIntyre, K. (2010). Charting new territory: Assessing the online frontier of student ratings of instruction. *Assessment & Evaluation in Higher Education, 35*(1), 101–115. https://doi.org/10.1080/02602930802618336

Uttl, B., White, C.A. &. Wong Gonzalez, D. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54,* 22-44.

Waschull, S. B. (2001). The online delivery of psychology courses: Attrition, performance, and evaluation. *Computers in Teaching, 28*, 143–147. https://doi.org/10.1207%2FS15328023TOP2802_15

Weschke, B. & Canipe, S. (2010). The faculty evaluation process: The first step in fostering professional development in an online university. *Journal of Teaching & Learning, 7*(1), 45-58.

Wheeler, R. W., Wright, L. & Frost, C. (2005). The effects of receiving positive or negative primes on an instructor's evaluation. *North American Journal of Psychology, 7*(1), 151-160.

Whitworth, J., Price, B., & Randall, C. (2002). Factors that affect college of business student opinion of teaching and learning. *Journal of Education for Business, 77*, 282–289. https://doi.org/10.1080/08832320209599677

Young, S. & Duncan, H. (2014). Online and face-to-face teaching: How do student ratings differ? *Journal of Online Learning and Teaching, 10*(1), 70-79. Retrieved from http://jolt.merlot.org/vol10no1/young_0314.pdf