Collegiate Aviation Review International

Volume 38 | Issue 2 Peer-Reviewed Article #3

8-4-2020

Bias and Trends in Student Evaluations in Online Higher Education Settings

Cheryl Lynn Marcham
Embry-Riddle Aeronautical University

Ann Marie Ade
Embry-Riddle Aeronautical University

Patti Clark
Embry-Riddle Aeronautical University

James Marion
Embry-Riddle Aeronautical University

End-of course evaluations have been frequently used to assess teaching effectiveness and influence critical decisions about faculty contract renewal, future course assignment, tenure and promotion in higher education. This quantitative study sought to determine whether there are differences in student perceptions of faculty performance based on gender or faculty status (full-time vs. adjunct) in an online higher education environment. It also sought to answer these questions: 1) Do adjunct faculty tend to grade more leniently than full time faculty, and as such, do adjunct faculty receive higher evaluation ratings than full time faculty, who may be more stringent in grading? 2) Do student evaluation scores differ depending on the course being evaluated? 3) Does gender or faculty status impact student response rates? Survey responses from a total of 683 sections associated with 24 courses were analyzed from the March 2018 to January 2019 timeframe. Due to the broad range of class sizes and differences between faculty characteristics, the variances for each comparison sample were observed to be significantly different using Levene's test for equal variances. Thus, the Mann-Whitney test for two variables and the Kruskal-Wallis test for evaluation of significant difference between more than two variables were used on the data. While other literature and personal anecdotes may indicate that gender bias exists, this study did not indicate that gender bias is occurring in online higher education courses taught for the time period studied, suggesting gender neutrality.

Recommended Citation:
Marcham, C.L., Ade, A.M., Clark, P. & Marion J. (2020). Bias and Trends in Student Evaluations in Online Higher Education Settings. *Collegiate Aviation Review International*, 38(2), 34-50. Retrieved from http://ojs.library.okstate.edu/osu/index.php/CARI/article/view/8036/7417

The use of student evaluations is ubiquitous at institutions of higher education, and often, important decisions are made based on student evaluation data. For example, administrators use teaching evaluations for annual review, promotion, tenure, and reappointment decisions. Department heads may consider results from evaluations to decide whether to keep a course or course content in the curriculum or to change it. Because the results from student evaluations can have such high stakes, it is important that we understand the limitations of any potential bias that might occur from a variety of sources or conditions, or bias towards a particular category of recipient.

Gender Influences

Previous research has illustrated that gender differences have historically been prevalent in student end-of-course and instructor evaluations in traditional brick and mortar settings. In 1989, a study of 9,005 student evaluations found that female professors, overall, had lower ratings than males for teacher effectiveness, academic competence, sensitivity to student needs, and overall performance; these differences held even while controlling for a number of variables such as students' sex, GPA, expected grade, discipline, and course size (Andersen & Miller, 1997; Sidanius & Crane, 1989). In 1991, Statham, Richardson, and Cook reported that there were differences in gender expectations for university instructors, and as a result, differences in how instructors were evaluated. For instance, the more classroom time a woman professor spent in presenting material, the lower her likability ratings, but the reverse was true for the male professors (Statham, Richardson, & Cook, 1991). Checking students' understanding and soliciting their input also enhanced the women's competence ratings but had a strong negative impact on both competence and likability ratings for men (Statham, Richardson, & Cook, 1991).

A gender bias can still be found in more current student evaluations of traditional university classroom instructors. A study of 19,952 student evaluations of university faculty at the School of Business and Economics of Maastricht University in the Netherlands over the period 2009-2013 found that, on average, female instructors systematically received a score 37 percentage points lower than male instructors, a bias primarily driven by male students' evaluations (Mengel, Sauermann, & Zolitz, 2018). Student evaluation data from the University of Oregon consisting of over 36,000 data sets collected from 2010 to 2016 were evaluated by Ancell and Wu (2017), who found that female instructors received course evaluation scores, on average, 0.0578 points lower than male instructors.

In some cases, the difference in ratings between male and female instructors has been attributed to students having different expectations for male versus female instructors. As described earlier Statham, Richardson, and Cook (1991) showed that historically, and in a traditional classroom setting, differences in gender expectations resulted in differences in how instructors were evaluated. This difference is consistent with the role congruity theory (Eagly & Karau, 2002) where students may expect female instructors to behave according to female gender stereotypes and male instructors to behave according to male gender stereotypes, but still evaluate overall teaching competence for all instructors according to the characteristics of the

stereotypical male professor (Boring, 2017; Kierstead, D'Agostino, & Dill., 1988; Basow, Phelan, & Capostosto, 2006; MacNell, Driscoll, & Hunt 2015). These gender stereotypes are still found in current studies. Boring evaluated 20,197 student evaluation scores over five academic years from traditional classroom courses and found that male students gave significantly higher overall satisfaction scores to male professors than to female professors. Boring also found that, in this study, a male professor's expected excellent overall satisfaction score was approximately 20% higher than a female professor's expected excellent overall satisfaction score, even though students performed equally well on final exams whether their professor was a man or a woman, suggesting no difference in actual teaching effectiveness. Thus, Boring posited that differences in teaching skills were not driving the gender differences in evaluations. In 2019, in a study of more than 523,000 student evaluations with more than 3,100 instructors, Fan et al. found that male students gave lower scores to female instructors regardless of the cultural backgrounds of either student or instructor. Clearly, there is an abundance of information indicating that gender bias against female instructors in student evaluations may still be occurring, at least in the traditional classroom setting.

Course subject may also have an impact on overall evaluation scores. Beran and Violato (2005) found that evaluations for courses in social sciences received significantly higher ratings than courses in natural sciences. Uttl and Smibert (2017) found that evaluations for quantitative classes like those in math received much lower average class summary ratings than non-quantitative classes such as those in English, history, or psychology. Related to this issue are studies that have shown that gender bias in student evaluations may also be more significant for some fields of study than others (Rosen, 2017). Fan et al. (2019) found that where there are larger proportions of female teachers, such as in the Arts and Social Sciences, there is less gender bias in student evaluations of teaching. Conversely, in technical and scientific areas of study, more gender bias may be prevalent.

With the increasing number of university courses moving to an online environment, one question that arises is whether gender bias becomes less predominant in a distributed environment. Online higher education has been promoted as an *equalizer* that breaks down the access barrier, and not only provides access for students from diverse cultures, but from diverse situations and economies all over the world (Black, Bissessar, & Boolaky, 2019). Cohen and Ellis, in 2008, posited that asynchronous learning networks (ALN) offered the potential to create a gender neutral communication environment. However, Mitchell and Martin (2018) report that when comparing evaluations for instructors teaching identical online courses, the language students used in evaluating a male professor was significantly different than the language used in evaluating a female instructor, and the students gave higher ordinal scores in the teaching evaluation to a male instructor than to a female instructor, even for questions specific to the course, not to the instructor. MacNell et al. (2015) found similar results in that students rated the instructors they perceived to be female lower than those they perceived to be male, regardless of teaching quality or actual gender of the instructor. These differences in student ratings were not a result of gendered behavior on the part of the instructors, but of actual bias and differing expectations on the part of the students. For example, when male and female instructors posted grades after two days as a male, this was considered by students to be a 4.35 out of 5 level of promptness, but when the same two instructors posted grades within the same time frame as a female, it was considered to be a 3.55 out of 5 level of promptness (Macnell et al., 2015).

However, both of these studies have limited sample sizes, as one involved only two instructors during a single term and the other involved only 43 students in a single 5-week summer class at a large public institution with over 20,000 students. Boring, Ottoboni, and Stark performed nonparametric statistical evaluation of over 23,00 evaluations from both the Boring study (originally published in 2015) and the Macnell, Driscoll & Hunt study, and confirmed bias against female instructors "by an amount that is large and statistically significant" (Boring et al., 2016b, para. 1). These researchers found that instructors whom students believed were male received significantly higher average ratings than those whom students believed were female (Boring et al., 2016b).

Grade Influences

Another issue of concern is when institutions focus on student evaluation data to make faculty review, promotion, tenure, and reappointment decisions; many instructors may choose to please the students with reduced scrutiny of assignments and higher grades to ensure high evaluation rates. Johnson (2003) argued that the onset of the importance given to student evaluations has brought about rampant grade inflation, as professors realized they could achieve better evaluation scores through easier grading. Stroebe (2016) continued this work, showing that that while the grade point average at colleges and universities has increased for decades, the amount of time students devote to their studies has continuously decreased. Stroebe (2016) argues that this grade inflation is:

> …encouraged by the practice of university administrators to base important personnel decisions on student evaluations of teaching. Grading leniency creates strong incentives for instructors to teach in ways that would result in good student evaluations. Because many instructors believe that the average student prefers courses that are entertaining, require little work, and result in high grades, they feel under pressure to conform to those expectations. (p. 800)

A 2016 survey of faculty members by the American Association of University Professors, revealed that 67 percent concurred that student evaluations put upward pressure on grading practices (Doerer, 2019). Ancell and Wu (2017) found that for each one point in increase in the GPA of a class led to between a 0.182 and 0.319 point increase in the instructor's evaluation score. Braga, Paccagnella, and Pellizzari (2014) found that teachers of classes that are associated with higher grades received better evaluations from their students. Numerous additional researchers have confirmed that instructor ratings have been found to correlate with student grades in the course (Adams & Umbach, 2012; Crumbley and Reichelt, 2009; Isely and Singh, 2005; Marsh 2007; Carrell & West, 2010; Krautmann & Sander, 1999; Weinberg, Hashimoto, & Fleisher, 2009; Boring et al., 2016b). Connected to this correlation is the concern that numerous studies that show that adjunct faculty in higher education institutions assign higher grades than full-time faculty (Reynolds, 2015; Cavanaugh, 2006; Kezim, Pariseau, & Quinn, 2005; Lippmann, Bulanda, & Wagenaar, 2009; Sonner, 2000). In fact, Boring et al. (2016a) state that the evaluation process contributes to grade inflation.

Limitations of the Student Evaluation Process

Student evaluations are often given a high priority even though several studies show that there is no direct correlation between student evaluations and teaching effectiveness or student learning. Linse (2017) published guidelines for the use and interpretation of student ratings data. In these guidelines, Linse emphasizes that student ratings are student perception data, not faculty evaluations, and that student ratings are not measures of student learning. Doerer (2019) opines that often, students are treated as customers, and their evaluations are more a metric of student satisfaction, not academic progress. Boring et al.'s (2016b) statistical analyses of more than 23,000 evaluations of 379 instructors by 4,423 students concluded that the association between student evaluations and teaching effectiveness was weak and not statistically significant. To quote Flaherty on the issue, students' teaching evaluations, "measure students' gender biases better than they measure the instructor's teaching effectiveness" (2016, para. 1). Boring et al. (2016a) argue that the evaluations are not strongly associated with learning outcomes, and as such, evaluating ratings are "at best, weakly associated with student performance" (para. 5).

Canadian researchers conducted a meta-analysis of 97 studies that revealed that students do not learn more from professors with higher student evaluation ratings, and such ratings are unrelated to student learning. Further, research by Braga, Paccagnella, and Pellizzari (2014) found that teachers who were more effective in promoting future performance receive worse evaluations from their students, indicating that evaluation scores are not related to teaching effectiveness. In fact, a 2016 meta-analysis of 51 articles containing 97 multi-section studies on student evaluations of teaching (SET) concluded that:

> Despite more than 75 years of sustained effort, there is presently no evidence supporting the widespread belief that students learn more from professors who receive higher SET ratings. If anything, the latest large sample studies show that students who were taught by highly rated professors in prerequisites perform more poorly in follow up courses. (Uttl, White, & Gonzalez, 2017, p. 40)

Because of the potential for bias, and because there is not a documentable connection between student evaluations and learning, or between student evaluations and teaching effectiveness, several institutions have abandoned or restructured the student evaluation process. In Canada, the Ryerson University Faculty Association argued that because of well-documented bias in student evaluations, they shouldn't be used for personnel decisions (Doerer, 2019). In August, 2018, Ryerson University was ordered by an arbitrator to amend the faculty collective bargaining agreement to ensure that faculty course survey results are not used to measure teaching effectiveness for promotion or tenure (*Ryerson University v. Ryerson Faculty Association*, 2018). In September, 2018, The University of Southern California Academic Senate concluded that since "research on student evaluations show that results are not correlated with learning outcomes or other valid measures of teaching effectiveness," and since these evaluations are "prone to systematic bias against women and…faculty of color," that there was a "need for a more meaningful review of teaching than student evaluations provide" (University of Southern California Academic Senate, 2018, para. 4-5). In March 2019, the University of Oregon Office of the Provost posted that it was working with the University Senate to revise the teaching evaluation system because:

> Recent research suggests that student ratings may not accurately reflect the quality of teaching due to biases and other factors. The University of Oregon's own assessment of student course evaluation ratings have corroborated these findings. The Association of American Universities (AAU) and other universities around the globe from University of Colorado, Boulder to University College London, England have argued that it is time for universities' practices regarding teaching excellence and evaluation to align with their policies. As such, the University of Oregon seeks to develop a holistic new teaching evaluation system that does more than simply replace problematic evaluation instruments so that we can help the UO community more effectively define, develop, evaluate, and reward teaching excellence. (para. 1-2)

After performing a comprehensive meta-analysis of 97 studies, Uttl, White, and Gonzalez (2016) suggested that because there was little to no significant correlation found between evaluation rating and learning, "institutions focused on student learning and career success may want to abandon SET ratings as a measure of faculty's teaching effectiveness" (para.1).

Therefore, given the current reliance on end-of-course evaluations to assess faculty teaching effectiveness, contract renewal, tenure, and promotion decisions, an assessment of potential bias in student evaluations for faculty at a regionally accredited online university was undertaken. This study sought to determine whether there are differences in the student perceptions of faculty performance based on gender or faculty status (full-time vs. adjunct). This study also sought to evaluate such questions as:

1. Do adjunct faculty tend to grade more leniently than full time faculty, and as such, do adjunct faculty receive higher evaluation ratings than full time faculty, who may be more stringent in grading?
2. Do student evaluation scores differ depending on the course being evaluated (i.e., if a course is poorly designed or particularly difficult, will that result in overall lower instructor evaluation scores, regardless of the instructor presenting the course)?
3. Does gender or faculty status impact student response rates?

The overall purpose was to identify potential bias that may affect future course, promotion or tenure decisions, based in part on current end-of-course survey responses, and whether there are any trends that can predict evaluation results. Given the nature of the focused curriculum (aviation/aerospace) and the predominance of male faculty and students at this university and within the target industry, any biases toward female faculty, or towards full time faculty who will not succumb to grade inflation pressure, may harm the potential of female or full time faculty to progress through the ranks of the university.

Methodology

The online campus for this study provides courses that are structured such that a master course outline and a master course template are provided to both full time and adjunct faculty assigned to teaching the course. Instructors are advised that no changes are to be made to the

course template, assignments, syllabus, or rubric. Therefore, the material presented, the manner in which it is presented, the assignments and assessments, as well as the grading structure are all consistent between instructors. Instructors are, however, encouraged to supplement the online course, and are expected to post personal biographical information, participate in weekly discussion boards, and regularly post announcements to engage the students.

A total of 683 sections associated with 24 courses taught in the online campus were selected from historical class records from the period of March 2018 to January 2019. Courses selected were those that were frequently taught by multiple instructors, had not been updated or changed during the study period, and were from a range of technical and general courses, including math, economics, aviation, English, research, and occupational safety topics. Student end-of-course survey responses, which are not required to be completed in order to obtain a final grade or any other service from the university, were collected for these course sections. By the very design of the end-of-course survey process, no personally identifiable data is collected about the student respondents. Grade distributions for each section of the course offered during the time frame as well as the data relating to the gender and employment status of the faculty member were collected and coded by the Office of Institutional Research to protect the identities of all participants, both faculty and students in the selected sections of courses for analysis. The categories of data collected from each course included the following:

- Course number and title
- Full-time/part-time instructor status
- Instructor gender
- End-of-course evaluation question response rates
- Class grade point average (GPA) per course
- End-of-course evaluation question scores for the following questions:
    - The instructor exhibited expertise in the course subject matter
    - My overall impression of the instructor is positive
    - The instructor provided meaningful and timely feedback on my assignments and progress

End-of-course evaluation scores are on a Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). All data collected can be available to other researchers upon request.

Based on the data collected, the following research questions were evaluated:

1. Is there a significant difference in GPA between courses?
2. Is there a significant difference in class GPA between male and female instructors for all classes?
3. Is there a significant difference in class GPA between full time and part time instructors for all classes?
4. Is there a significant difference in end-of-course evaluation question scores between male and female instructors?
5. Is there a significant difference in end-of-course evaluation question scores between full time and part time instructors?
6. Is there a significant difference in end-of-course evaluation question scores between

courses?

7. Is there a relationship between course GPA outcomes and student evaluation response scores?
8. Is there a significant difference in end-of-course evaluation question response rates between faculty genders?
9. Is there a significant difference in end-of-course evaluation response rates between full-time and part-time faculty?

All research questions except for research question 7 involved tests of significant differences for one or more variables. The raw data for each research question was evaluated for equality of variances using Levene's test for equal variances. Due to broad range of class sizes and differences between the number of male versus female and full time versus part time faculty, the variances for each comparison sample were observed to be significantly different. All tests of significance therefore used the Mann-Whitney test for two variables, and the Kruskal-Wallis test for evaluation of significant difference between more than two variables.

Table 1
Faculty Composition and Class Size Information

Faculty Status	**Faculty Gender**	**Mean Class Size**	**Class Size Standard Deviation**
576 Part Time	499 Male	20	8
107 Full Time	184 Female		

Results & Discussion

For the research questions addressing differences in GPA between courses, between male and female instructors, and between full and part time instructors, no significant difference was found between any of these variables and the overall GPA of the class. See Table 2 for test statistic values. Of particular note, this finding indicates that grade inflation is not occurring with part time instructors compared to full time instructors, at least for the courses evaluated.

For the research questions addressing end-of-course evaluation scores, again, no difference was found between male and female instructors or between full-time and part-time instructors with one exception (see Table 2 for test statistic values). For the end-of-course question, "The instructor provided meaningful and timely feedback on my assignments and progress," no significant difference was found between full-time and part-time instructors at the 95% level, however, the .0617 *p*-value is within 1.2% of the accepted $p = .05$ level. This finding suggests that response to this end-of-course question does exhibit some difference between full-time and part-time instructors. Overall, the mean score for full-time instructors was found to be 4.299 whereas the mean score for part-time instructors was 4.440.

Table 2
Man-Whitney Test Results

Research Question	**Test Statistic**	***p* Value**	**Results**
GPA differences between courses	18.78	.2055	No significant difference was found in course GPA
GPA differences between male and female instructors	.3322	.7937	No significant difference was found in course GPA between male and female instructors
GPA differences between full-time and part-time instructors	.6715	.5019	No significant difference was found in course GPA between full-time and part-time instructors
End of course evaluation score differences between male and female instructors ("The instructor exhibited expertise in the course subject matter.")	-.0791	.9370	No significant difference was found in course evaluation scores between male and female instructors
End of course evaluation score differences between male and female instructors ("My overall impression of the instructor is positive")	.0158	.9874	No significant difference was found in course evaluation scores between male and female instructors
End of course evaluation score differences between male and female instructors ("The instructor provided meaningful and timely feedback on my assignments and progress.")	.9333	.3506	No significant difference was found in course evaluation scores between male and female instructors
End of course evaluation score differences between full-time and part-time instructors ("The instructor exhibited expertise in the course subject matter.")	-1.051	.2933	No significant difference was found in course evaluation scores between full-time and part-time instructors
End of course evaluation score differences between full-time and part-time instructors ("My overall impression of the instructor is positive.")	-.8466	.3972	No significant difference was found in course evaluation scores between full-time and part-time instructors
End of course evaluation score differences between full-time and part-time instructors ("The instructor provided meaningful and timely feedback on my assignments and progress.")	-1.8685	.0617	No significant difference was found in course evaluation scores between full-time and part-time instructors at the 95% level, however, the .0617 p value is within 1.2% of the accepted p=.05 level. This finding suggests that response to this end-of-course question does exhibit some difference between full-time and part-time instructors.
Differences in response rates related to faculty gender	.9125	.3615	No significant difference found in course response rates based upon faculty gender.
Differences in response rates related to instructor employment status (full-time or adjunct	-3.228	<.01	There is a significant difference found in course response rates based upon faculty employment status.

Difference in end-of-course response scores were further evaluated to determine whether there was any significant difference in course response scores between courses identified as technical/scientific versus those classified a non-technical/arts and social science. While previous research has indicated that gender bias may be more prevalent in scientific and technical areas of study (Fan et al., 2019), this bias was not found to be the case with the

evaluations studied at this university.

Response rates were also evaluated. There was no significant difference found in course response rates based upon faculty gender or between courses (Kruskal-Wallis, 25.068, $p = .296$), but there was significant difference found in course response rates based upon faculty employment status. Response rates for part-time instructors was higher than for full-time instructors, but that may be a function of sample size, with 575 part-time instructors analyzed compared to only 107 full-time instructors.

To evaluate whether there is a difference in course evaluation scores between courses, a Kruskal-Wallis test was performed since there were more than two variables. Using this test, a test statistic of 101.57 with a p-value <.01 was found. Therefore, a significant difference in evaluation scores was found between courses. Some courses had an overall mean evaluation score of as low as 2.60, whereas the highest mean score for one particular course was 3.58. This may support the hypothesis that student evaluations differ depending on the course (i.e., if a course is poorly designed or particularly difficult, that may result in overall lower evaluation scores, regardless of the instructor presenting the course). Looking at the mean scores by course may be a valuable tool for administration to identify courses that may need attention, and may also be useful in explaining why individual instructors may receive low evaluations when teaching certain courses.

When evaluating whether a relationship exists between course GPA outcomes and student evaluation response scores, a correlation analysis was performed. A positive yet relatively weak correlation was found for evaluation questions "The instructor provided meaningful and timely feedback on my assignments and progress" and "The instructor exhibited expertise in the course subject matter" (both $r = .22$, $p < .01$). However, there is a stronger association ($r = .27$, $p < .01$) for the question "My overall impression of the instructor is positive." It was observed in this analysis that positive impressions increase with higher grades.

Limitations

One important impact on data integrity is the impact of nonresponse rates, which can increase the potential for error and weaken the quality of data and their results (Groves et al., 2004; Groves & Couper 1998). In the age of data-driven decision-making, it is imperative to collect and use responses representative of the whole population, but many universities fail in obtaining high response rates, particularly those from online evaluation processes (Adams & Umbach, 2012). Adams and Umbach (2012) report that in most cases, survey nonreseponse rates are not random. Bacon, Johnson, and Stewart (2016) confirmed that when response rates are low, high-scoring teachers are rated much more favorably, and low-scoring teachers are rated much less favorably, most likely because those students that do respond have a strong opinion, but the would-be scores from those who did not respond were not present to balance out the overall score. As nonresponse rates increase, the likelihood increases that the opinions of those who did not complete the survey differ from those who did, thus the data in these student surveys are not always representative of the whole population (Adams & Umbach, 2012). Multiple studies report that response rates for online student evaluations can initially average near 60%, but often drop off to the 30 to 40 percentile range (Avery, Bryant, Mathios, Kang, & Bell, 2006; Nulty,

2008; Sax, Gilmartin, & Bryant, 2003). Chapman and Joines (2017) have recommended minimum response rates for class sizes over ten, under liberal conditions (10% sampling error, 80% confidence level), a minimum response rate of 70% is recommended (Chapman & Joines, 2017). While some of the online classes evaluated for this university could have class sizes of under 10, the overall mean response rate for the courses evaluated for this study was 77%.

It is recognized that the larger the number of statistical tests performed, the greater the risk of Type I errors, or false positive results (Andrade, 2019; Armstrong, 2014). Methods such as the Bonferroni or Hochberg corrections are available (Andrade, 2019; Armstrong, 2014), but were not used in these evaluations. The study results produced very few positives thereby reducing the need for tests of false positives.

Conclusions & Recommendations

While the historic literature and personal anecdotal experiences of individual instructors may indicate that gender bias can occur, the analysis of over 683 data points does not indicate that gender bias is occurring in courses taught online or hybrid environment at this university for the time period studied. To recap the study parameters, a total of 683 sections associated with 24 courses taught in the online campus were selected for the period of March 2018 to January 2019. The courses were chosen to fit multiple parameters such as frequently taught by multiple instructors, had not been updated or changed during the study period, and were from a range of technical and general courses, including math, economics, aviation, English, research, and occupational safety topics. The data utilized was gleaned from the course section student end-of-course survey responses and GPA differences as detailed in Table 2. What should be an obvious point is that a lot of data was compiled and analyzed for this study. Through meticulous examination of the data, the authors concluded that no evidence of gender bias was evident in the end of course survey responses or differences in GPAs. Conclusions allow us to be introspective and draw inferences from the results. The conclusions were unexpected, and the results are certainly contrary to the majority of previous studies conducted on traditional classroom environments. However, the results corroborate the earlier theorization of Cohen and Ellis (2008) that ALN offer the potential to create a gender neutral communication environment and we conclude from this study that online and hybrid modalities muted gender bias in the data examined.

Beyond the lack of gender bias detected in the data, one relationship that should be pointed out is the relationship between course GPA outcomes and student evaluation response scores. The weak yet positive correlation found in evaluation questions “The instructor provided meaningful and timely feedback on my assignments and progress” and “The instructor exhibited expertise in the course subject matter” was not a surprise to the authors. When considered with the weak but stronger association for the question “My overall impression of the instructor is positive” the inference can be drawn that a student will report a positive impression of an instructor when a higher GPA in the course is achieved. Again, while not unexpected and a belief often articulated by instructors, the conclusion is troubling from a perspective that the student may perceive the instructor is the basis for the high grade rather than the grade was earned through the student’s efforts in the course. This particular issue is perhaps a conundrum that has existed as long as instructors have scored student submissions and awarded final course grades.

While the research questions evaluating bias for this study were not supported by the evidence, that fact is perhaps the most encouraging and enlightening aspect of the research. As a community of higher education institutions, we are embracing online teaching technology at an ever increasing rate with new institutions entering the market daily. The Education Department's National Center for Education Statistics reported that in 2017 of all students in postsecondary courses students in mixed online and in person courses accounted for 17.6% of enrollments and students exclusively in online courses stood at 15.4% of all enrollments (Lederman, 2018). As the demand for online and hybrid learning grows, as has occurred exponentially in 2020 as a result of the COVID-19 pandemic, so do the opportunities to make the learning environment truly gender neutral. We all strive for an environment where both faculty and students are accepted and valued and not viewed through a gender bias lens.

This research establishes an important foundation for other studies in the evolving online education environment. Online learning is persistent and the numbers support the acceptance of the modality by students even in the advent of declining postsecondary enrollments (Lederman, 2018). The authors suggest future studies be undertaken that examine student gender bias in the online environment. Does gender neutrality extend to the actual students in an online or hybrid learning environment course? Other research threads should be considered that delve deeper into the association of student course GPA to positive impressions of the instructor. The weak yet positive correlations discovered in this study indicate a more in depth inquiry into a student's perceptions of earned versus awarded grades is warranted. Additionally, the student evaluation process should be vetted further to determine whether it is a useful or outdated tool particularly for online learning environments. Should teaching effectiveness be evaluated by the data and not the student as in an online learning environment? A plethora of data resides in each course to evaluate not only faculty teaching effectiveness, but other factors that influence student evaluations today such as time in course to GPA, timeliness of grading and assignment learning outcome alignment to name a few aspects.

As noted earlier, the value of this research lies in what was absent in the data and not what was present. Bias of any type marginalizes individuals and in a learning environment it can be toxic to effectiveness of the faculty member. Moving forward, let's continue to foster this gender neutrality in online environments and take additional measures to ensure students are judged impartially as well.

References

Adams, M. J., & Umbach, P. D. (2012). Nonresponse and online student evaluations of teaching: Understanding the influence of salience, fatigue, and academic environments. *Research in Higher Education,* 53(5), 576-591.

Ancell, K., & Wu, E. (2017). Teaching, learning, and achievement: Are course evaluations valid measures of instructional quality at the University of Oregon? Retrieved from https://provost.uoregon.edu/files/course_evaluations_wu_ancell.pdf

Andersen, K., & Miller, E. D. (1997). Gender and student evaluations of teaching. *PS: Political Science and Politics*, 30(2), 216-219. doi:10.2307/420499

Andrade, C. (2019). Multiple testing and protection against a type 1 (false positive) error using the Bonferroni and Hochberg corrections. *Indian Journal of Psychological Medicine*, 41(1), 99-100. doi:10.4103/IJPSYM.IJPSYM_499_18

Armstrong, R. A. (2014). When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, 34(5), 502-508. doi:10.1111/opo.12131

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic SETs: Does an online delivery system influence student evaluations? *Journal of Economic Education*, 37, 21–37.

Bacon, D. R., Johnson, C. J., & Stewart, K. A. (2016). Nonresponse bias in student evaluations of teaching. *Marketing Education Review*, 26(2), 93-104. doi:10.1080/10528008.2016.1166442

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(1), 25–35.

Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment & Evaluation in Higher Education*, 30(6), 593-601.

Black, D., Bissessar, C., & Boolaky, M. (2019). Online education as an opportunity equalizer: The changing canvas of online education. *Interchange*, 50, 423-443. doi:10.1007/s10780-019-09358-0

Boring, A. (2015). Working paper: Gender biases in student evaluations of teaching. Retrieved from https://www.ofce.sciences-po.fr/pdf/dtravail/WP2015-13.pdf

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

Boring, A., Ottoboni, K., & Stark, P. B. (2016a). Student evaluations of teaching are not only unreliable, they are significantly biased against female instructors [Blog post]. Retrieved from https://blogs.lse.ac.uk/impactofsocialsciences/2016/02/04/student-evaluations-of-teaching-gender-bias/

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

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

Carrell, S. E., & West, J. E. (2010). Does professor quality matter?: Evidence from random assignment of students to professors. *Journal of Political Economy*, 118(3), 409-432. doi:10.1086/653808

Cavanaugh, J. K. (2006). What did you get? A faculty grade comparison. *Quality Assurance in Education: An International Perspective*, 14(2), 179-186.

Chapman, D. D., & Joines, J.A. (2017). Strategies for increasing response rates for online end-of-course evaluations. *International Journal of Teaching and Learning in Higher Education*, 29(1), 47-60.

Cohen, M. S., & Ellis, T. J. (2008, October). *The asynchronous learning environment (ALN) as a gender-neutral communication environment*. Paper presented at the 38th ASEE/IEEE Frontiers in Education Conference, Saratoga Springs, NY. doi:10.1109/FIE.2008.4720279

Crumbley, D. L., & Reichelt, K. J. (2009). Teaching effectiveness, impression management, and dysfunctional behavior: Student evaluation of teaching control data. *Quality Assurance in Education*, 17(4), 377–392.

Doerer, K. (2019). Colleges are getting smarter about student evaluations. Here's how. *The Chronicle of Higher Education*, 65(18), A8.

Eagly, A., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109(3), 573–598.

Fan Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019) Gender and cultural bias in student evaluations: Why representation matters. *PLoS ONE* 14(2): e0209749. https://doi.org/10.1371/journal.pone.0209749

Flaherty, C. (2016, January 11). Bias against female instructors. *Inside Higher Ed.* Retrieved from https://www.insidehighered.com/news/2016/01/11/new-analysis-offers-more-evidence-against-student-evaluations-teaching

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. New York: Wiley

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley.

Isely, P., & Singh, H. (2005). Do higher grades lead to favorable student evaluations? *Journal of Economic Education*, 36(1), 29–42.

Johnson, V. (2003). *Grade inflation: A crisis in college education*. New York: Springer

Kezim, B., Pariseau, S. E., & Quinn, F. (2005). Is grade inflation related to faculty status? *Journal of Education for Business*, 80(6), 358-363.

Kierstead, D., D'Agostino, P., & Dill, H. (1988). Sex role stereotyping of college professors: Bias in students' ratings of instructors. *Journal of Educational Psychology*, 80(3), 342–344.

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

Lederman, D. (2018, November 7). Online education ascends. *Inside Higher Ed*. Retrieved from https://www.insidehighered.com/digital-learning/article/2018/11/07/new-data-online-enrollments-grow-and-share-overall-enrollment

Linse, A. R. (2017). Interpreting and using student ratings data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation*, 54, 94-106.

Lippmann, S., Bulanda, R. E., & Wagenaar, T. C. (2009). Student entitlement. *College Teaching*, 57(4), 197-204.

Marsh, H.W. (2007). Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases and usefulness. In R.P Perry & J.C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319-383). Dordrecht: Springer. doi:10.1007/1-4020-5742-3_9

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(4), 291. doi:10.1007/s10755-014-9313-4

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

Mitchell, K., & Martin, J. (2018). Gender bias in student evaluations. *Political Science & Politics*, 51(3), 648-652. doi:10.1017/S104909651800001X

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, 33(3), 301-314.

Reynolds, D. (2015). Variability of passing grades in undergraduate nursing education programs in New York State. *Nursing Education Perspectives*, 36(4), 232-236. doi:10.5480/13-1235

Rosen, A. S. (2017). Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: A large-scale study of RateMyProfessors.com data. *Assessment & Evaluation in Higher Education*, 43(1), 31-14. doi:10.1080/02602938.2016.1276155

Ryerson University v. Ryerson Faculty Association, CanLII 58446 (2018)

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. *Research in Higher Education*, 44(4), 409–432.

Sidanius, J. & Crane, M. (1989). Job evaluation and gender: The case of university faculty. *Journal of Applied Social Psychology*, 19, 174-97.

Sonner, B. S. (2000). "A" is for 'Adjunct': Examining grade inflation in higher education. *Journal of Education for Business*, 76(1), 5-8.

Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science*, 11(6), 800-816. doi:10.1177/1745691616650284

Statham, A. Richardson, L., & Cook, J. A. (1991). *Gender and university teaching: A negotiated difference*. Albany, NY: State University of New York Press.

Uttl, B. & Smibert, D. (2017). Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career. *PeerJ*, 5, e3299.

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

University of Oregon Office of the Provost. (2019, March). Revising UO's teaching evaluations. Retrieved from https://provost.uoregon.edu/revising-uos-teaching-evaluations

University of Southern California Academic Senate. (2018, September 20). Teaching evaluations update. Retrieved from https://academicsenate.usc.edu/teaching-evaluations-update/

Weinberg, B. A., Hashimoto, M., & Fleisher, B. M. (2009). Evaluating teaching in higher education. *The Journal of Economic Education*, 40(3), 227-261. doi:10.3200/JECE.40.3.227-261