COMMERCIAL SOFTWARE FOR TEACHER COURSE EVALUATIONS

ABSTRACT

Student evaluations of courses have been researched over the years to see whether or not they actually measure teaching effectiveness, what role they should play in tenure and promotion, and which characteristics of courses lead to better outcomes. In this paper, we propose to focus on another dimension: whether a commercial product can help improve faculty and course evaluation numbers. We compare four year of measurements using an in-house product with four years of measurements using a commercial product. We find that the longer commercial for, with its in-depth analysis did not lead to significantly higher instructor or course outcomes.

INTRODUCTION

With colleges, programs, and courses under pressure to justify their cost and existence, course evaluations play an ever more important role to many parties [6][13][15][19]. Administrators need to be able to evaluate instructor performance, to show the usefulness of courses to students, and to evaluate the well-being of programs offered by the school. Faculty want to be assured that the topics they cover and the way they approach their classes are indeed of benefit to the students. Students want to enroll in those courses that other students have ranked as most beneficial to them. Characteristics of courses and their effect on student evaluations have been a topic of interest for many researchers [7][16][20] and other have focused on whether or not these student evaluations really measure teaching quality [2][3][10][11][14][17][21].

Two major shifts have taken place in recent years in the world of course evaluations. The first shift is from giving written evaluations in-class to an on-line model of data collection. Benefits of this approach are numerous: the data can be read more easily, all data is in one place, standardized forms are used across classes enabling more direct comparisons, and analysis is quickly done by computer, to name a few. This transition has been studied extensively by many authors. The major drawback of the on-line process is that fewer students complete the form. This leads faculty to protest their use as measures of goodness. This shift has been studied extensively in the literature. As expected, the studies reported that online evaluations had a significantly lower response rate than in-class evaluations [8][9][19][4][5]. Strategies have been suggested to increase student response rates [8][5]. The studies showed that there was no significant change in instructor and course ratings between online and in-class evaluations [4][12]. An extensive literature list addressing the various issues related to online and in-class evaluations is given in [1].

The second shift is the move from school-designed evaluations to commercial products. Some examples of commercial products, to name just a few, are Evaluation Kit, IDEA, Explorance, and Snap Surveys. Unlike most evaluations that have been developed in-house, commercial products have been designed by experts in the fields of education, statistics, and methodology. In addition to providing tested questions relevant to teacher performance, these products have extensive guidelines for faculty about ways they can improve their classes. Through analysis of selected question groupings, the evaluation summaries give faculty guidance on what they do well and specifically where they need improvement. Not much seems to have been studied about this second shift. Thus, the question is, for the expense of these commercial products, are they making any difference in the teaching performances of faculty?

This paper is a case study of that question. We would like to investigate whether or not the commercial evaluation tool method has had a beneficial effect on teacher performance in our school. To do this, we look at the average instructor score and the average course score on end-of-course evaluations done with an in-house product (we will abbreviate this as IHP), and with a commercial product (hereafter called COMM). We are a large mid-western private university that has switched from an in-house evaluation to an external commercial evaluation product. We have data on teacher scores in one school for an eight-year period. During the first four years of this period, a 20 question survey that had been developed in-house was administered online. In the last four years, a 40 question survey from a commercial provider was administered, also online. Note that some of the commercial products give administrators an option to choose a shorter version of the questionnaire with fewer questions and allow the instructor to add custom questions. In this comparison, we will look at instructors and courses over eight years and track the changes in performance over that time period. Our overall question was whether or not the commercial product with its extensive analysis and feedback yielded improved scores on student evaluations of faculty teaching or on course performance.

DATA

The data for this study consists of course evaluation data for eight years. We will call these years IHP1 through IHP4 and COMM1 through COMM4. Faculty included fell into one of three groups: (1) they taught all 8 years, (2) they taught the 4 years prior to the change, but not after, or (3) they taught for all 4 years after the switch to the new evaluation tool, but not before. This includes a total of 84 individual faculty: 49 in all years, 16 evaluated only with the IHP, and 19 evaluated only with the COMM. Courses were also broken up into three parts. We have courses taught the entire 8 year period, courses taught only before the switch of evaluation tools, and new courses occurring only in the 4 years following the tool change. We have 2,885 course sections (1,427 course sections with IHP and 1,458 with COMM) representing 63 continuously taught courses, 12 courses unique to the IHP, and 12 courses evaluated only with COMM. We track the movement of evaluation averages for instructor and for course across these eight years. For the first four of these years, our school used an in-house online evaluation tool that consisted of 20 questions. The in-house tool contained, at the end, two summative questions about the overall rating of the instructor and of the course. For the last four years, we used an online commercial evaluation tool with 40 questions. This tool also included similar questions about the overall rating of the instructor and of the course. Each term, the analysis of student responses was made available to each course instructor shortly after the ending of the term. In this paper, we are focusing only on the two questions common to both questionnaires: the overall rating of the course and the overall rating of the instructor.

The in-house evaluation tool analysis reported, for each section of a course, the raw response counts plus an average to each question. Students were also given space for general written comments. These were copied and returned to faculty. No further analysis of the scores was done. The commercial tool is an online system of course evaluations that not only allows students to rate many aspects of courses, but also gives an in-depth analysis of the average responses to faculty to help them understand how they might improve their courses in the future. The student receives a link to an on-line survey with 40 multiple choice questions that can be rated on the Likert-scale from 1 (strongly disagree, or very poor, depending on the format of the statement) to 5 (strongly agree, or excellent). They also have space to write commentary about anything they wish. Prior to student access to the questionnaire, faculty are asked to select their main learning objectives for the course from a group of questions. The summary that is presented to the teacher includes an evaluation of each question in the survey and also how the course was rated with respect to the teacher-identified learning objectives. The multiple-choice responses to questions

are reported to the faculty in their raw form (for example 4 5s, 16 4s, etc.) and as an overall average (4.20). The multiple choice questions are further grouped into categories with suggested actions for the faculty to improve their scores in these areas. Qualitative written responses are also reported, but not summarized.

The switch from an in-house instrument to the one produced commercially had several assumed benefits. First, it would standardize the evaluations across all courses and all schools, enabling easier comparisons for the university across subject areas. Second, the processing would not be done in-house, freeing our IT division from this labor-intensive task. Lastly, the presentation of results would help faculty to improve their teaching by furnishing an in-depth analysis of each class. It is this third purpose that we wish to focus on in this paper. Did the commercial tool analysis improve faculty performance in the classroom? We will address this by looking at the four questions listed here. We also, to maintain privacy, select an average base score to which the scores in each year are compared. Thus, all scores in years IHP1 through COMM4 are differences from this base score.

- For faculty who taught all 8 years, is the instructor average for the fourth year of the commercial product (COMM4) significantly better than the initial Base instructor average?
- For new faculty who taught only the last 4 years, is the instructor average for year COMM4 significantly better than the initial Base instructor average?
- For courses that were taught all 8 years, is the course average for year COMM4 significantly better than the initial Base course average?
- For new courses that were taught only the last 4 years, is the course average for year COMM4 significantly better than the initial Base course average?

RESULTS

Our data included a score for each instructor in the years using the IHP and the years using the COMM evaluation. We had corresponding evaluation scores for the courses. That is, both the IHP and the COMM asked the student to give an overall score to the instructor, and to give an overall score to the course. It is these two questions that we are analyzing. We first report the changes in average scores over the first year (IHP1), for each year, for the instructors and courses. These values are reported in Table 1 and Table 2. In these tables, the years using the in-house product are on the left, followed by the commercial product values on the right. To maintain confidentiality, scores are shown as differences from a base score.

For faculty and courses that were in the entire 8-year span, we want to compare year IHP1 to year COMM4. That is, after several years of exposure to the new commercial tool with its extensive analysis, we would hope the faculty had significantly improved over their original scores with the in-house tool. Similarly for courses, in looking at courses that we have taught during the entire 8-year period (for example, introductory statistics), we will compare the scores from years IHP1 and COMM4. The comparison scores, for both the faculty and course, for the entire 8-year period are reported in Table 1. For faculty who joined when the new evaluation form was beginning to be used, but who were not part of the faculty during the IHP years, we compare their performances to the Base instructor value. Similarly, for new courses introduced later, we will compare the evaluation scores of each year with the Base course value. These comparison scores for COMM1, COMM2, COMM3, and COMM4 are reported in Table 2. We also reported the scores comparing the performances of new and experienced faculty in Table 3. In this table, the scores shown are the differences of the average scores given to new faculty (ie. Faculty only here during the COMM period) and those given to faculty who were here the entire eight years. Similarly for courses, we compare courses that are new in the COMM period with the classic courses that have continued over the entire 8 years. In this table, a positive value indicates that the new faculty/course got higher scores and a negative value means the continuing faculty/courses did better.

Table 1. Changes in averages for faculty and courses over 8 years, comparison years in bold.

	IHP1 - Base	IHP2 - Base	IHP3 - Base	IHP4 - Base	COMM 1-Base	COMM2 - Base	COMM3 - Base	COMM4 - Base
Faculty in All 8 Years	0.000	-0.058	-0.028	-0.009	-0.120	-0.178	-0.185	-0.180
Faculty Only Before	-0.521	-0.602	-0.699	-0.842				
Courses in All Years	0.000	-0.022	0.021	0.004	-0.037	-0.063	-0.058	-0.088
Courses Only Before	-0.084	-0.083	-0.206	-0.183				

Table 2. Changes in averages for faculty and courses after COMM

					COMM1-Base	COMM2-Base	COMM3-Base	COMM4-Base
Faculty Only After					-0.237	-0.262	-0.405	-0.254
Courses Only After					0.159	-0.038	-0.136	-0.124

Table 3. Comparison: (New – Experienced) Faculty and (New – Continuing) Courses

					COMM1	COMM2	COMM3	COMM4
Faculty: (New – Experienced)					-0.117	-0.084	-0.221	-0.074
Course: (New – Continuing)					0.159	-0.012	-0.115	-0.073

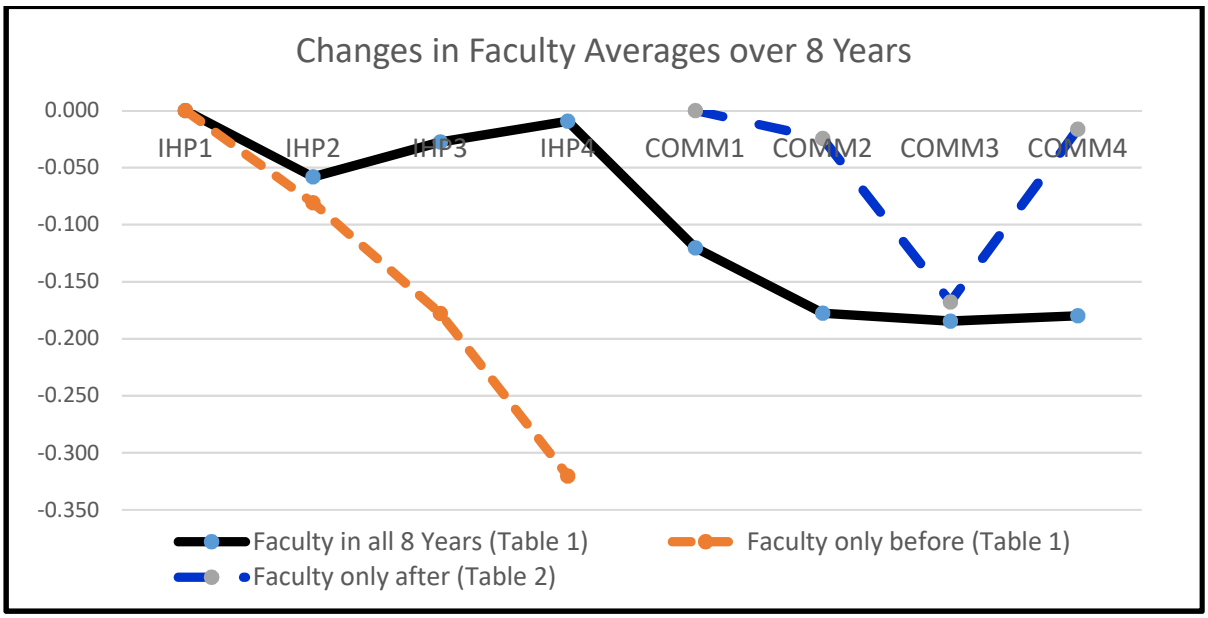
In Table 1, we see that, in each of the cases, the beginning score was higher than the ending score. For the faculty who taught during the entire 8-year stretch, we see a drop of 0.18 from IHP1 to COMM4. If we look at the four years using the in-house product, there is not a lot of variation in these overall faculty scores.

The first year using the commercial product we see the largest drop of 0.11 from the year IHP4 to COMM1. Since we are analyzing an identical question across all years, it is surprising that this drop is larger than any previous year. But we would expect the following years to show increased improvement year-by-year if it was simply that the faculty were not familiar with the new survey. Instead, we see that the scores in following years never showed an increase over the Base score. Could this be due to the length of the commercial product? Why didn't all the analysis and recommendations of the new product help?

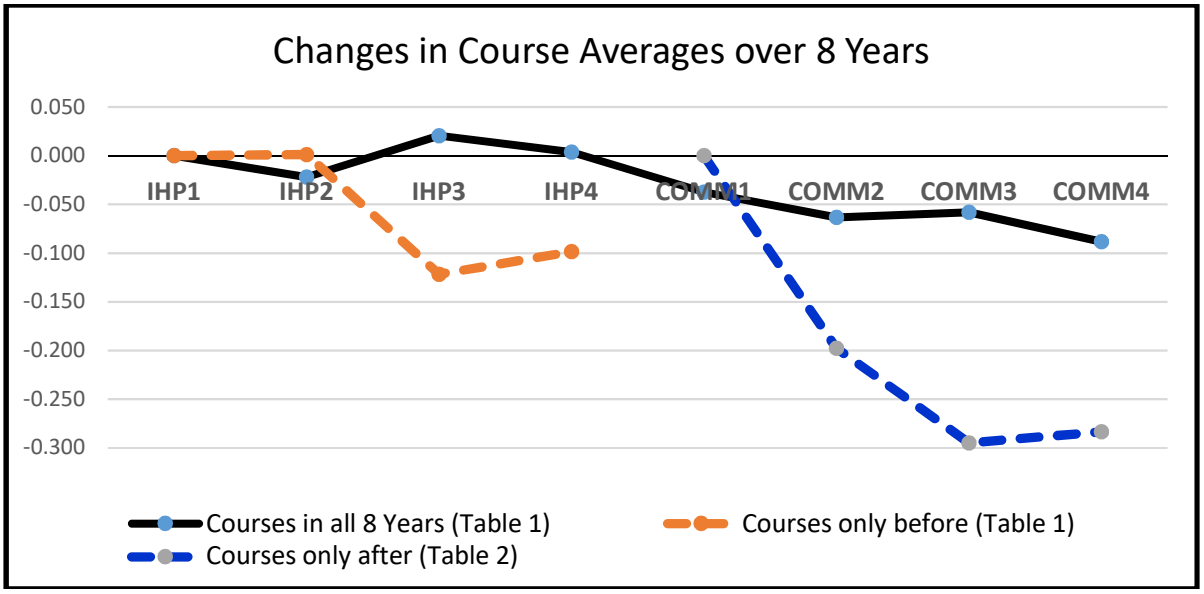
For courses, we see a similar pattern. For the set of courses that were taught during the entire 8 years, we see a drop of 0.037 from IHP1 to COMM1. The years under the in-house product had similar values. In the transition from IHP4 to COMM1, we see a slight drop of 0.04. This drop continued, and the value in the last year, COMM4, was lower than in any other year. For the new courses that were introduced in the year of COMM1, we see the highest score of all for courses. Then a drop of additional 0.05 in COMM4 over COMM1. Thus, the extensive analysis and suggestions offered by the commercial product did not have the positive effect we had hoped for.

Graphs 1, 2, and 3 below allow us to visually see the movement of each of these group averages over time. If we compare the faculty values that are in Table 1, we see that no final value is higher than the beginning value in any of the three groups. But some other interesting patterns also appear here. For the group of faculty who left the school (either through retirement or to pursue other jobs), we see that their overall averages decreased each year over the period by a total of 0.321. We also see that, under the extensive feedback of the commercial product, faculty scores drifted lower each year. They never recovered the value they had in any year while using the in-house product.

Graph 1. Changes in Faculty Averages Over 8 Years



Graph 2. Changes in Course Averages Over 8 Years

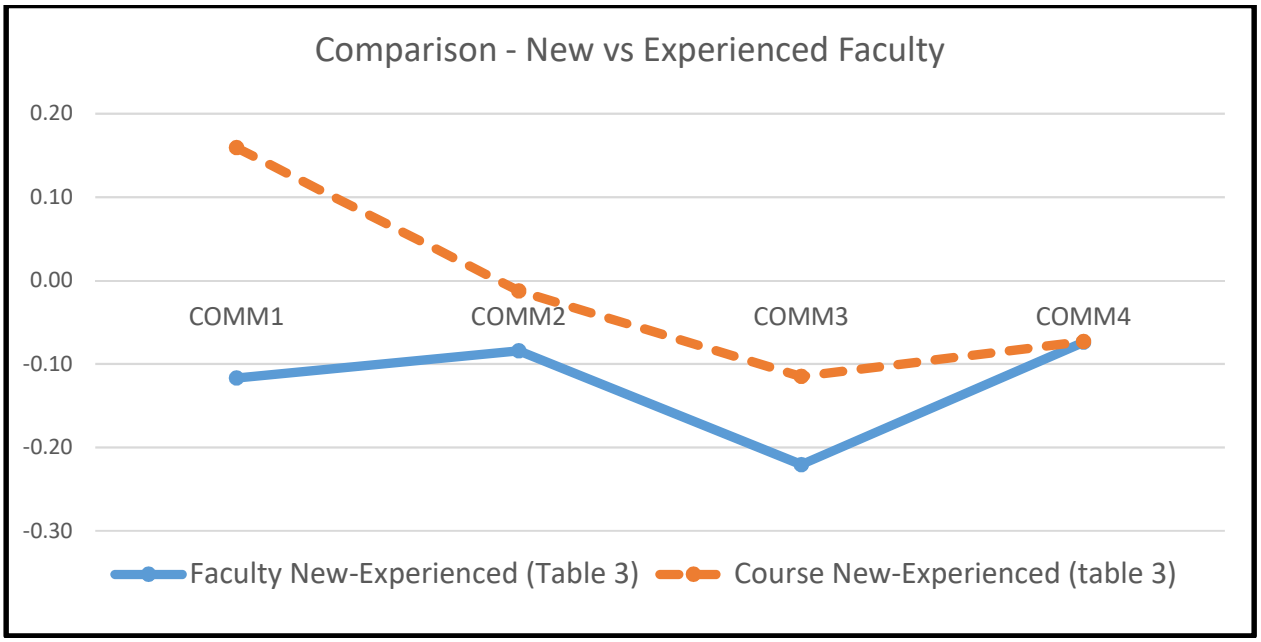


New courses experienced the highest drop in the evaluations (Graph 2). In the courses that were phased out, we see that their evaluation by students was below the average group of courses. This might be the

perception of older topics as of less use to a current student. We see also that newly introduced courses have a much higher score in their first year than in later years. This might be due to greater engagement from both faculty and students when looking at new topics. Since these are not the main questions that are the focus of this paper, we leave their analysis for another time.

Graph 3 looks at the difference scores from Table 3. Here, we have the calculated score of (new – existing) for both faculty and courses. When these scores are positive, new faculty/courses are evaluated more highly, when they are negative, the continuing faculty/courses received higher numbers. New faculty joining the school were slightly lower by 0.117 than experienced faculty in their first year and remained slightly lower (by 0.073) than the experienced faculty in their last year (Graph 3). New courses started off higher, but dropped off over time.

Graph 3. Comparison – New vs Experienced Faculty



Next, we turn our attention to a statistical comparison of the numbers in Tables 1 and 2. We consider the four questions stated at the end of the DATA section. We repeat them here is an abbreviated format:

1. For faculty who taught all 8 years, was the COMM4 score significantly greater than the base score?
2. For new faculty who taught only the last 4 years, is their COMM4 score significantly greater than the base score?
3. For courses that were taught all 8 years, is the COMM4 score significantly greater the base score?
4. For new courses that were taught only the last 4 years, is the COMM4 score significantly greater than the base score?

The results of a paired t-Test for two sample means for these are shown in Table 4.

Table 4. Statistical Tests

Question	t-Statistics	Critical Value	Conclusion
1	-2.788	1.677	for continuing instructors, COMM4 is not statistically better
2	-0.159	1.734	for new instructors, COMM4 is not statistically better
3	-1.429	1.670	for continuing courses, COMM4 is not statistically better
4	-1.481	1.796	for new courses, COMM4 is not statistically better

As we see from the values here, in no case was the value in COMM4, the fourth year using the commercial product statistically greater that the comparison value.

CONCLUSION

We looked at eight years of data on student evaluation of instructors and courses. This data was divided into two parts. In the first four years, we used an in-house evaluation tool that consisted of about 20 multiple choice questions with room for open responses at the end. The last four years used a purchased commercial tool that was longer (about 45 multiple-choice questions followed by room for open ended responses). The in-house product returned question counts and averages to the instructor, plus a copy of the open ended questions. The commercial tool contained these plus an in-depth analysis of what the instructor should do to improve.
As we can see from the graphs and t-tests from the previous section, neither the student's evaluations of the instructors or the courses increased as a result of using the commercial product to evaluate the courses. Student participation varied some over the years. It began at 52% and ended at 43%.

We can hypothesize a number of reasons that are often voiced when discussing student evaluations of teaching: the number of students responding to evaluations is down; only students who have complaints fill out the forms; the cost of higher education increases expectations for the classroom; or students expect to just be given good grades. Anecdotally, we hear from faculty that the questionnaire is too long and frustrates students, that the summary given to the faculty is too long and involved to be helpful (typically 4 or 5 pages of analysis).

But, for whatever reason, we do find that the hopes we had for the commercial form were not fulfilled. Neither the instructor nor the course evaluation numbers increased in the way we had envisioned.

REFERENCES

[1] Barre, B. Student Ratings of Instruction: A Literature Review, 2015, https://cte.rice.edu/blogarchive/2015/02/01/studentratings.

[2] Bassett, J., A. Cleveland, D. Acorn, M. Nix and T. Snyder. Are they paying attention? Students' lack of motivation and attention potentially threaten the utility of course evaluations, *Assessment & Evaluation in Higher Education*, 2017, 42(3), 431-442, DOI: 10.1080/02602938.2015.1119801.

[3] Braga, M., M. Paccagnella, M. Pellizzari. Evaluating students' evaluations of professors, *Economics of Education Review*, 2014, 41, 71-88.

[4] Capa-Aydin, Y. Student evaluation of instruction: comparison between in-class and online methods, *Assessment & Evaluation in Higher Education*, 2016, 41(1), 112-126, doi: 10.1080/02602938.2014.987106.

[5] Crews, T.B, D. F. Curtis. Online Course Evaluations: Faculty Perspective and Strategies for Improved Response Rates, *Assessment & Evaluation in Higher Education*, 2011, 36(7), 865-878, doi: 10.1080/02602938.2018.1532491.

[6] Ellis, T. Completing the cycle: meaningful course evaluations, *33rd Annual Frontiers in Education, 2013, doi*: 10.1109/FIE.2003.1263363.

[7] Gorry, D. The impact of grade ceilings on student grades and course evaluations: Evidence from a policy change, *Economics of Education Review*, 2017, 56, 133-140.

[8] Guder, F., M. Malliaris. Online and Paper Course Evaluations. *American Journal of Business Education*, 2010, 3(2), 131-137.

[9] Guder, F., M. Malliaris. Online Course Evaluations Response Rates. *American Journal of Business Education*, 2013, 6(3), 333-337.

[10] Hoefer, P., Yurkiewicz, J., & Byrne, J. C. The association between students' evaluation of teaching and grades, *Decision Sciences Journal of Innovative Education,* 2012, *10*, 447–459. doi:10.1111/j.1540-4609.2012.00345.x.

[11] Hornstein, H. A. Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance, *Cogent Education*, 2017, 4(1).

[12] Kuch, F., R. M. Roberts. Electronic in-class course evaluations and future directions, *Assessment & Evaluation in Higher Education*, 2019, 44(5), 726-731, doi: 10.1080/02602938.2018.1532491.

[13] Mitchell, K., & Martin, J. Gender Bias in Student Evaluations. *PS: Political Science & Politics*, 2018, 51(3), 648-652. doi:10.1017/S104909651800001X.

[14] Oermann, M. Student Evaluations of Teaching: There Is More to Course Evaluations Than Student Ratings, *Nurse Educator*, 2017, 42(2), p. 55-56, doi: 10.1097/NNE.0000000000000366.

[15] Rahman, S. A Web-Based Course and Instructor Online Evaluation System, 2015 Fifth International Conference on e-Learning, 18-20 Oct. 2015, Bahrain, doi: 10.1109/ECONF.2015.24.

[16] Robinson, S. Mixing It Up: The Impact of Resequencing Topics in an Undergraduate Introductory Statistics Course, 2019, PRIMUS, doi: 10.1080/10511970.2019.1600178.

[17] Stark, P. and R. Freishtat. An Evaluation of Course Evaluations, *ScienceOpen Research*, 2014, https://www.scienceopen.com/document/vid/42e6aae5-246b-4900-8015- dc99b467b6e4?0, doi: 10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1.

[19] Stowell, J. R., W. Addison, J. L. Smith. Comparison of online and classroom-based student evaluations of instruction, *Assessment & Evaluation in Higher Education*, 2012, 37(4), 565-473, doi: 10.1080/02602938.2010.545869.

[19] Supiano, B. and D. Berrett. Everyone Hates Course Evaluations, *The Chronicle of Higher Education*, Nov. 30, 2017, https://www.chronicle.com/article/Everyone-Hates-Course/241929.

[20] To, W.M. & M. Tang. Computer-based course evaluation: an extended technology acceptance model, *Educational Studies*, 2019, 45(2), 131-144, doi: 10.1080/03055698.2018.1443797

[21] Venkatesh, V., A. Croteau, J. Rabah. Perceptions of Effectiveness of Instructional Uses of Technology in Higher Education in an Era of Web 2.0, 2014, 47th Hawaii International Conference on System Sciences, Waikoloa, HI, doi: 10.1109/HICSS.2014.22.