Croatian Journal of Education
Vol.21; No.3/2019, pages: 743-775
Original research paper
Paper submitted: 13th April 2018
Paper accepted: 28th November 2018
https://doi.org/10.15516/cje.v21i3.3177

The Validity of Student Evaluation of Teaching: Is There a Gender Bias?

Milica Maričić, Aleksandar Đoković and Veljko Jeremić
University of Belgrade, Faculty of Organizational Sciences, Department of Operational Research and Statistics

Abstract

Student evaluation of teaching (SET) has steadily, but surely, become an important assessment tool in higher education. Although the SET provides feedback on the student level of satisfaction with the course and the lecturer, the validity of its results has been questioned. Following extensive studies, the factor which is believed to distort the SET results is the lecturer's gender. In this paper, Potthoff analysis is employed to additionally explore whether there is any gender bias in the SET. Namely, this analysis has been used with great success to compare the linear regression models across groups. Herein, we aimed to model the overall lecturer impression with independent variables related to teaching, communication skills and grading, and to compare the models between genders. The obtained results reveal that the gender bias exists in certain cases in the observed SET. We believe that our research might provide additional insights into the interesting topic of gender bias in the SET.

Key words: *educational data mining; gender bias; higher education; Potthoff analysis; student evaluation of teaching.*

Introduction

Since the 1970s, the use of student evaluation of teaching (SET) has expanded dramatically (Kogan, Schoenfeld-Tacher, & Hellyer, 2010) and, slowly but surely, it has become an almost universally accepted survey for obtaining feedback from the main university stakeholders, students (Zabaleta, 2007). Nowadays, student evaluation questionnaires around the world are quite similar when it comes to their structure and questions. Namely, most of them are based on the five- or seven-point Likert scales and ask students to express their agreement or disagreement with statements on the course materials, the teaching style, and the lecturer (Kogan et al., 2010). After

the surveys are completed by hand or online, the responses are usually summarised across instructors, departments, and faculties. Although the process at a glance appears simple, in reality, it is far more complex and susceptible to the impact of internal and external factors. The three major issues which the SET encounters are the mostly ordinal type of data, the subjective nature of questions, and the multi-faced structure of student satisfaction (Simonacci & Gallo, 2017).

As the SET can provide valuable information to the university administration and government bodies, it has become mandatory in many countries around the world (OECD, 2009). Namely, the SET results are used in both governmental and private auditing, and accreditation procedures (Johnson, 2000). The results of SET can also have an impact on the lecturers' advancement in the academia. In some countries, these results are an important part of the academic CV and make a difference when it comes to hiring and promotion (Maricic, Djokovic, & Jeremic, 2016).

We can observe that the SET results can have multiple impacts on the teaching performance, academic staff, and accreditation process. Therefore, the SET is a valuable source of information on teaching and student satisfaction with the course or the lecturer (Chen & Hoshower, 2003). Having that in mind, the SET should be created, distributed, and interpreted with caution and precision. Nevertheless, the question which was raised by many experts in the field of education assessment and lecturers is the SET validity (Zhao & Gallant, 2012). In other words, lecturers question students' competence to grade teachers, classes, and course syllabuses, and they also believe that the SET does not have one widely accepted definition of the concept "effective teaching" (Johnson, 2000). On the other hand, experts have tackled the SET on both national and international level (La Rocca, Parrella, Primerano, Sulis, & Vitale, 2017). The main concerns are related to the SET structure and questions, along with the validity of students' answers, which are susceptible to other factors rather than teaching (Maricic et al., 2016; Zhao & Gallant, 2012).

Extensive research has been conducted to identify the factors which might have an effect on the students' evaluation of lecturers. For example, the respondents' gender may affect the results. Basow (2000) has shown that the gender of the student has an impact on the choice of the "best" teacher and the SET scores. Another interesting research by Centra (2003) analysed the relation between the SET results and the workload and the course difficulty. He provided evidence that the courses seen as difficult were always rated lower. Also, Shevlin and associates (2000) have shown that the SET results are easily influenced by the teacher's charisma. As we can see, the factors which may distort the SET results can be divided into three groups: student-related, course-related, and teacher-related (Pounder, 2007).

The aim of this study is to explore the influence of the teacher-related factor on the SET scores – the gender of the lecturer. Experts in the field of higher education, sociology, and anthropology believe that gender is a significant factor in teacher

evaluations, which should be further explored (Basow, Phelan, & Capotosto, 2006; MacNell, Driscoll, & Hunt, 2015). Although the SET literature has significantly increased in the last several decades, we attempted to improve on the existing studies. The main hypothesis is that different teaching aspects have different impacts on the overall lecturer's score depending on his/her gender. First, we aim to investigate whether there is a difference in the impact of teaching skills and style on the overall SET score by the lecturer's gender and by subject. And second, we present the factors influencing the overall SET score of a lecturer by gender at the level of a particular higher educational institution.

The paper begins with a literature review on the presence of gender bias in SET. It then introduces the statistical methodology applied to examine the possible presence of gender bias. The central part refers to the conducted SET and the obtained research results. In the final two parts, the potential limitations of our study are discussed, further research directions are proposed, and the obtained results are elaborated.

Gender Bias in the Student Evaluation of Teaching (SET)

To better understand the observed phenomenon, we provide a brief literature review on the gender bias in the student evaluation of teaching. The topics in our focus are the gender of the teacher, the positive and negative traits of the chosen "best" teachers, and the SET scores on the teaching style, grades, and lecture content.

In the research on the factors that influence the SET results, gender is one of the most researched (Liu, 2012). Studies related to gender bias proliferated in the 1980s and 1990s (Basow, 1995; Bennett, 1982). However, throughout the years, the results have been inconclusive, and the presence of gender bias remained an unresolved question. On the one hand, gender bias has been reported (Basow, 1995; Maricic et al., 2016), while other studies have stated that gender bias does not exist, and that male and female lecturers are evaluated along similar dimensions (Aleamoni, 1999; Arbuckle & Williams, 2003). Therefore, the conflicting results make it unclear whether gender bias exists in the SET scores and whether the lecturer's gender plays a complex and multi-faceted role in the SET scores (Basow et al., 2006).

Sprague and Massoni (2005) have conducted an interesting research in which they asked male and female students to describe their "best" and "worst" teacher. When it comes to the top five traits of "best" male and female teachers, the results show that, for male teachers, it is important who they *are* as a person, while for female teachers it is important what they *do* during classes. This is an important discovery as it shows that different teaching approaches are desirable for male and female teachers, and that the same effort is valued differently. Namely, they reveal that there are significant differences in the teaching style, which students value in male and female teachers.

One of the factors that significantly affect the SET scores is the teaching style. Namely, research by Boring (2015) has showed that students have different approaches

when it comes to evaluating the teaching styles of male and female lecturers. Her extensive study shows that male lecturers are awarded higher scores for the teaching dimensions which are not time-consuming, such as animation and presentation skills. However, the situation is entirely different for female teachers who are praised for the teaching dimensions that demand more time, such as course preparation, class organisation, feedback, and consultations.

An interesting research was conducted by Abel and Meltzer (2007), in which they explored whether there would be differences in the evaluation of a male and female lecturer after presenting the identical lecture on a gender-related topic. They showed that the students rated the female professor and her lecture as more sexist and gave her lower overall ratings even though the male professor had had the same lecture. This result shows that the lecturer's gender is taken into account when grading the comprehensibility of the lecture and the lecture itself. However, this study might have a limitation as the subject of the lecture was gender-related, which might have emphasised the gender stereotypes among students.

Double standards are applied when evaluating the lecturer's objectivity in grading students' assessments. Female lecturers are more often judged, seen as incompetent, and punished with the lower SET scores than their male colleagues when they give the same low grades to students, according to Sinclair and Kunda (2000). They conducted an experiment in which the students had to evaluate their male and female lecturers after receiving the test scores. The results went in favour of the male lecturers/experts even when they gave lower grades and negative feedback. In other words, female lecturers are expected to give higher grades and provide students with positive feedback and comments (Boring, 2015).

If there is a difference in the students' evaluations of teachers based on the gender stereotypes, then male and female teachers have to teach, communicate, and mark students differently to obtain the high SET scores. The university administration should be aware of the presence of gender bias in the SET scores, and should bear in mind that male and female lecturers are under a burden to meet the students' gender expectations (Sprague & Massoni, 2005). Following the presented literature, a study was conducted to examine whether gender has an impact on the importance of teaching, communication, and assessment on the overall SET score of the lecturer.

Methodology

Application of Multivariate Analyses and Data Mining Techniques in the Assessment of SET

Linear regression modelling is often employed to uncover the relationship between specific questionnaire items and overall teaching score (Attanasio & Capursi, 2011). Namely, although questionnaire items are usually on the five- or seven-point Likert scales, the linear regression has been applied with great success. For example, Nowell,

Gale and Handley (2010) used various ordinal and scale variables to create two models and to explain the average of five lecturer's characteristics and overall lecturer's evaluation. In a more recent study, Jiang et al. (2016) have employed data mining techniques and multivariate regression to perform a longitudinal study into an undergraduate course at a large engineering faculty.

Besides the linear regression, other multivariate analyses have been employed to explore the SET scores. For example, Remedios and Lieberman (2008) used the principal component analysis (PCA) to assess the factors that influence the SET scores. Ho, Watkins and Kelly (2001) performed the one-way multivariate analysis of variance (MANOVA) to evaluate the effects of a staff development programme on students' learning approaches.

Therefore, we can observe that various multivariate analyses could be employed on the SET scores. However, the linear regression deemed a good method for several reasons. First, as mentioned above, it can be used to evaluate and explore the SET which is mostly based on the five-point Likert scale. The SET, scrutinised in this research (explained in detail in Section *Conducted research*), has such structure. Second, the research question was not posed to group variables and to explore the factors that influence the SET scores, but to explore and compare the impact of certain lecturers' traits. Finally, there is a statistical analysis, Potthoff analysis, which allows us to compare regression models between groups with much success.

Potthoff Analysis

Multiple books and research articles were aimed at introducing methods and tests to compare the coefficients from the ordinary least squares regression models (Howell, 2013; Potthoff, 1966). Potthoff analysis stands out as a simple way of comparing the linear regression models. Namely, so far it has been used with success in the fields of social psychology (Lawson & Lips, 2014), innovation management (Truong, Klink, Fort-Rioche, & Athaide, 2014), and others. The benefit of this analysis lies in the fact that it provides information on whether regression models, created on different groups with the use of the same variables, differ or not. In other words, it can occur that the linear regression model is statistically significant for one group, but not for the other. Also, different coefficients might be statistically significant depending on the group. Evidence that the idea of comparing linear regression models is still developing and is valuable for the academic community, especially Potthoff analysis, is to be found in the recently published SAS and SPSS code for the related tests (Weaver & Wuensch, 2013).

Essentially, Potthoff analysis consists of three tests: test of coincidence, test of intercepts, and test of parallelism. The test of coincidence aims to explore whether there is a difference between the intercept, the slope or both, between the observed groups (Wuensch, 2016). Further, the test of intercepts tests whether the intercepts are identical across groups. Finally, the test of parallelism should provide information

on whether the slopes significantly differ. For more details on the analysis, consult Wuensch (2016).

There are several benefits of applying Potthoff analysis to inspect the presence of gender bias. First, it answers whether a categorical variable, in this case, the lecturer's gender, moderates the relationship between continuous variables (Lawson & Lips, 2014). Namely, the analysis provides insights into whether slopes or intercepts or both caused a difference between the two regression equations made for each gender. Thus, it effectively compares the regression models. Secondly, the difference in the regression models and their slopes can be a valuable source of information for both lecturers and university administration. Finally, the analysis can be conducted on ordinal data (Lawson & Lips, 2014), which is most commonly collected through the SETs.

Conducted Research

SET – Participants, Procedure, SET Structure

Participants of the conducted study were undergraduate students of the University of Belgrade's Faculty of Organizational Sciences. All the surveyed students were enrolled in the full-time courses within one of the four study programmes available: Information Systems and Technologies (IS&T), Management, Operational Management, and Quality Management. Herein, for our analysis, we used the SET results from the mandatory spring semester. The analysed SET is a compulsory and unified survey, which the faculties of the University of Belgrade conduct twice a year, at the end of each semester. Students were given an opportunity to rate their teachers and teaching associates, whose classes they had attended. The SET was distributed during lectures in the second half of May in 2016. It was administered to students by a volunteer while the teacher/associate was in the amphitheatre/classroom. Afterwards, the survey results were imported using Blaise, and the statistical analysis was performed using SPSS 22. The final SET scores were shared with the teacher/associate after the end of semester; the best five teachers and five teaching associates were publicly commended in front of the Faculty Council.

The SET consists of three sections. The first section is devoted to the basic information about the subject and the lecturer: the study programme, the evaluated subject, the name of the evaluated teacher, and the date of evaluation. The second section aims to gather more information about the student and his/her grades. The items are related to the student's way of financing studies (scholarship or self-finance), his/her previous average grade, whether he/she previously attended the course, whether he/she regularly attended classes, and how many hours he/she weekly devoted to studying the subject. The last section differs when it comes to assessing teachers and associates. It consists of 11 teacher-related statements and 9 associate-related statements on which the students should express their agreement or disagreement on a five-point Likert scale in a range from 1 (strongly disagree) to 5 (strongly agree), including 0 (No answer). The statements used in this study are those related to the lecturer's assessment and they are presented in Appendix 1. The SET used to assess the teaching associates does not include item 4 "*Compliance of the lecture and the scope of the subject*" and item 7 "*Scope and quality of the suggested literature*".

Results

In our analysis, we aimed to create the linear regression models of "*Overall impression*" by using all the items related to the lecturer's evaluation. Our research was two-fold: first, on the level of particular subjects, and second, on the level of the Faculty of Organizational Sciences. The obtained models were later compared between genders using Potthoff analysis. The obtained results show that the differences exist on various occasions.

Out of 66 subjects, whose lectures were assessed for our analysis, we chose 6 subjects. We did not take into consideration the elective subjects, as the self-selection of students would complicate the analysis (Braga, Paccagnella, & Pellizzari, 2014). The chosen subjects had to be mandatory either for the whole generation or the whole study programme. Also, they had to have both male and female teachers and/or associates and more than 100 student assessments. It should also be remembered that one student could assess more than one of the observed subjects and lecturers if he/she had attended their classes. In total, we analysed 278 teacher evaluations (115 female and 163 male) and 1358 teaching associate evaluations (648 female and 710 male). Table 1 presents the chosen subjects, the type of lecturer, the number of student assessments, and Cronbach's alpha per subject per gender. Herein, we conducted Cronbach's alpha to test the consistency of the students' answers. It is interesting to see that most of the analysed SETs, except Data Basis for the male teaching associates, Decision Theory for the female teaching associates, and Financial Management and Accounting for the female teaching associates, have Cronbach's alpha above the cutoff 0.7 (Nunnally & Bernstein, 1994), meaning that most of the scales are reliable and that the answers were consistent.

Table 1

Chosen subjects, type of lecturer, number of assessments, and Cronbach's alpha per subject per gender of the assessed lecturer

Subject	Lecturer	Gender	N	Cronbach's alpha
Mathematics 2	Teacher	Female	50	0.733
		Male	67	0.806
Discrete Mathematical Structures	Teacher	Female	65	0.840
		Male	96	0.795
Introduction to Information Systems	Teaching associate	Female	235	0.737
		Male	179	0.834
Data Base	Teaching associate	Female	211	0.747
		Male	186	0.644
Decision Theory	Teaching associate	Female	127	0.668
		Male	231	0.882
Financial Management and Accounting	Teaching associate	Female	75	0.628
		Male	114	0.803

To explore whether there is the presence of the gender bias, we employed Potthoff analysis. The first analysis included the Mathematics 2 teachers. The test of coincidence revealed a significant effect, indicating that the regression line differs between the male and female teachers, $F(12,93)=4.556$, $p<0.01$. This result means that there was a difference in the two regression models. Therefore, we continued with the analysis to determine whether there was a difference in the intercept or in the slopes or in both of them.

Table 2

Regression models for the female and male Mathematics 2 teachers

Gender of the teacher	Item	B	S.E.	β	Sig
Female	Intercept	-0.877	0.891		0.331
	Q_1	0.305	0.105	0.236	0.006
	Q_2	0.023	0.050	0.042	0.650
	Q_3	-0.171	0.135	-0.111	0.213
	Q_4	-0.112	0.197	-0.053	0.572
	Q_5	0.541	0.080	0.605	0.000
	Q_6	0.033	0.108	0.027	0.763
	Q_7	0.410	0.073	0.479	0.000
	Q_8	-0.22	0.077	-0.287	0.007
	Q_9	0.081	0.084	0.086	0.338
	Q_{10}	0.495	0.114	0.446	0.000
	Q_{11}	-0.17	0.062	-0.260	0.009
Male	Intercept	-1.056	0.753		0.166
	Q_1	0.087	0.132	0.048	0.514
	Q_2	-0.066	0.035	-0.148	0.061
	Q_3	0.347	0.092	0.398	0.000
	Q_4	0.151	0.084	0.157	0.079
	Q_5	0.108	0.079	0.112	0.178
	Q_6	0.129	0.083	0.149	0.128
	Q_7	0.079	0.077	0.095	0.312
	Q_8	-0.008	0.082	-0.010	0.919
	Q_9	-0.062	0.089	-0.073	0.485
	Q_{10}	0.321	0.101	0.284	0.003
	Q_{11}	0.133	0.047	0.219	0.006

The second test, the test of intercepts, revealed that the intercepts in the regression did not significantly differ between genders, $F(1,93) = 0.02$, $p > 0.05$. Finally, the test of parallelism indicated that the slopes significantly differed, $F(11,93) = 4.935$, $p < 0.01$. The models for the female and male teachers are presented in Table 2. The obtained R^2 is 0.809 for the female teachers, and 0.747 for the male teachers, and both models are statistically significant ($F_{female}(11,38)=14.670$, $p<0.01$; $F_{male}(11,55)=14.759$, $p<0.01$). Taking a closer look at the regression models (Table 2), it can be observed that

different aspects of teaching are important, depending on the gender. For the female mathematicians the important traits are the *Regularity of classes* (t=2.911, p<0.01), *The teacher encourages the student to take part in the class, to critically think, and to be creative* (t=6.744, p<0.01), *the Scope and quality of the suggested literature* (t=5.578, p<0.01), *The teacher gives the students useful information for their future work* (t=-2.837, p<0.01), *The professionalism in the communication with students* (t=4.344, p<0.01), and *The objectivity and unbiasness in grading and evaluating the students' knowledge* (t=-2.737, p<0.01). This result shows that the "nurturing" traits are highly valued in the female teachers, which is in concordance with the result obtained by Sprague and Massoni (2005). On the other hand, when it comes to the model for the male mathematicians, the *Comprehensibility and presentation of the course syllabus* (t=3.790, p<0.01), the *Professionalism in the communication with students* (t=3.159, p<0.01), and the *Objectivity and unbiasness in grading and evaluating the students' knowledge* (t=2.860, p<0.01) are significant for the overall impression.

Table 3

Regression models for the female and male Discrete Mathematical Structures teachers

Gender of the teacher	Item	B	S.E.	β	Sig
Female	Intercept	4.453	0.298		0.000
	Q_2	0.009	0.019	0.038	0.628
	Q_3	0.023	0.054	0.044	0.665
	Q_4	-0.172	0.055	-0.341	0.003
	Q_5	0.187	0.066	0.364	0.007
	Q_6	0.196	0.083	0.355	0.023
	Q_7	0.033	0.061	0.064	0.587
	Q_8	0.376	0.086	0.669	0.000
	Q_9	-0.074	0.157	-0.087	0.642
	Q_{10}	-0.462	0.174	-0.540	0.010
	Q_{11}	-0.007	0.024	-0.026	0.764
Male	Intercept	-0.632	0.479		0.190
	Q_2	0.007	0.016	0.008	0.645
	Q_3	0.118	0.098	0.025	0.232
	Q_4	-0.054	0.049	-0.021	0.271
	Q_5	-0.194	0.101	-0.039	0.058
	Q_6	0.005	0.041	0.002	0.898
	Q_7	-0.081	0.061	-0.022	0.191
	Q_8	-0.016	0.051	-0.005	0.749
	Q_9	0.062	0.029	0.051	0.036
	Q_{10}	0.359	0.050	0.142	0.000
	Q_{11}	0.914	0.027	0.910	0.000

The next SET we modelled was the SET which assessed the teachers of the subject Discrete Mathematical Structures. This model differs from the previous one as variable *The classes are held regularly* had to be removed since all the teachers had the same

score. The analysis showed that the models significantly differed as $F(11,139)= 117.264$, $p<0.01$ and that there was a statistically significant difference between both intercepts ($F(1,139)=81.356$, $p<0.01$) and slopes ($F(10,139)=108.954$, $p<0.01$). The two obtained models are given in Table 3.

For the female teachers R2 is 0.754 and 0.982 for the male teachers, while both models are statistically significant ($F_{female}(10,54)=16.587$, $p<0.01$;$F_{male}(10,55)=458.960$, $p<0.01$). In the model for female teachers, the intercept is statistically significant ($t=14.941$, $p<0.01$), together with *Compliance of the lecture and the scope of the subject* ($t=-3.122$, $p<0.01$), *The teacher encourages the student to take part in the class, to critically think, and to be creative* ($t=2.806$, $p<0.01$), *Lectures assist students to master the course syllabus* ($t=2.346$, $p<0.05$), *The teacher gives students useful information for their future work* ($t=4.371$, $p<0.01$), and *Professionalism in the communication with students* ($t=-2.660$, $p<0.01$). Again, the result shows that the "nurturing" traits are highly valued in the female teachers. As in the male teachers of mathematics, the last two variables are statistically significant (Q_9: $t=2.126$, $p<0.05$; Q_{10}: $t=7.215$, $p<0.01$) alongside the variable *The teacher answers students' questions and takes into account the students' comments* ($t=33.512$, $p<0.01$).

Then we observed the SETs of the Introduction to Information Systems teaching associates. The test of coincidence showed that there was no difference between the two models ($F(10,93)=1.293$, $p>0.05$). As the two models did not differ, the test of intercepts and the test of parallelism were not conducted.

Next, the overall impression of the Data Base teaching associates was modelled. A significant difference between the two models was found ($F(10,377)=4.080$, $p< 0.01$), whereas that difference is in the slopes ($F(9,377)=4.452$, $p< 0.01$), not in the intercept ($F(1,377)=0.032$, $p> 0.05$). The obtained R2 is 0.602 for the female, and 0.460 for the male teaching associates, and both models are statistically significant ($F_{female}(9,201)=33.834$, $p<0.01$; $F_{male}(9,176)=16.600$, $p<0.01$). The two models are presented in Table 4. Taking a closer look at the model for the female teaching associates, it can be observed that, besides the intercept ($t=-2.153$, $p<0.05$), three variables are significant: *The ability to encourage students to take part in the class, to critically think, and to be creative* ($t=2.579$, $p<0.01$), *The professionalism in the communication with students* ($t=6.864$, $p<0.01$), and the *Objectivity and unbiasness in grading and evaluating the students' knowledge* ($t=2.730$, $p<0.01$). The model for the male teaching associates showed that variables *The comprehensibility and presentation of the course syllabus* ($t=2.382$, $p<0.05$). *The classes assist students to master the course syllabus* ($t=3.119$, $p<0.01$), and the *Objectivity and unbiasness in grading and evaluating the students' knowledge* ($t=2.246$, $p<0.05$) were statistically significant.

Further, the overall impression regarding the Decision Theory teaching associates was modelled as the base for the Discrete Mathematical Structures teachers, so the variable *The classes are held regularly* had to be removed again. No significant difference between the two models was found ($F(9,340)=1.874$, $p>0.05$). Finally, we assessed the

SETs for the associates teaching Financial Management and Accounting. Interestingly, there was no difference between the two models ($F(10,169)=1.396$, $p>0.05$).

Table 4

Regression coefficients of the models of female and male teaching associates for the subject Data Base

Gender of the Teaching Associate	Item	B	S.E.	β	Sig
Female	Intercept	-0.799	0.371		0.033
	Q_1	0.047	0.087	0.032	0.593
	Q_2	-0.002	0.015	-0.007	0.876
	Q_3	0.007	0.073	0.008	0.918
	Q_5	0.174	0.067	0.203	0.011
	Q_6	0.03	0.078	0.028	0.700
	Q_8	0.029	0.052	0.035	0.569
	Q_9	-0.003	0.070	-0.003	0.967
	Q_{10}	0.811	0.118	0.522	0.000
	Q_{11}	0.064	0.024	0.141	0.007
Male	Intercept	-1.123	1.592		0.482
	Q_1	0.541	0.346	0.096	0.120
	Q_2	0.008	0.014	0.033	0.575
	Q_3	0.194	0.082	0.205	0.018
	Q_5	0.1	0.053	0.143	0.059
	Q_6	0.285	0.091	0.269	0.002
	Q_8	0.077	0.050	0.116	0.128
	Q_9	-0.009	0.087	-0.008	0.915
	Q_{10}	-0.002	0.103	-0.001	0.985
	Q_{11}	0.044	0.019	0.135	0.026

To additionally explore whether there is gender bias in the SET conducted at the Faculty of Organizational Sciences, we performed the same analysis for teachers and teaching associates on the institutional level. We analysed 4162 teacher and 7056 teaching associate SETs. However, for this analysis, the items *The classes are held regularly* and *The consultations are held regularly* have been excluded due to a high frequency of score 5.

First, the teachers' overall impression was modelled. Potthoff analysis showed that there was a statistically significant difference between the two models as $F(10,4142) = 14.314$, $p<0.01$, that there was a difference in the intercept ($F(1,4142) = 5.447$, $p<0.01$), and in the slope ($F(9,4142) = 14.869$, $p<0.01$). The models for the female and male teachers are presented in Table 5. The obtained R^2 is 0.517 for the female and 0.477 for the male teachers. Both models are statistically significant ($F_{female}(9,2211)=263.006$, $p<0.01$; $F_{male}(9,1931) =195.867$, $p<0.01$). Interestingly, there are six traits that are important for both female and male teachers: Q3 ($t_{female}=3.151$, $p<0.01$; $t_{male}=4.772$, $p<0.01$), Q6 ($t_{female}=4.066$, $p<0.01$; $t_{male}=3.192$, $p<0.01$), Q7

(t_{female}=-5.260, p <0.01; t_{male}=-5.184, p <0.01), Q9 (t_{female}=17.053, p <0.01; t_{male}=12.888, p <0.01), Q10 (t_{female}=6.454, p <0.01; t_{male}=13.600, p <0.01), and Q11 (t_{female}=24.578, p <0.01; t_{male}=13.363, p <0.01). For the male teachers, besides the six traits, also *The compliance of the lecture and the scope of the subject* (t=2.671, p<0.01) and *The teacher gives students useful information for their future work* (t=4.419, p<0.01) are statistically significant. Another result which draws attention is that the quality of the suggested literature significantly decreases the overall SET score of both male and female teachers.

Table 5

Regression models for the female and male teachers of the Faculty of Organizational Sciences

Gender of the teacher	Item	B	S.E.	β	Sig
Female	Intercept	-0.268	0.132		0.043
	Q_3	0.102	0.032	0.064	0.002
	Q_4	0.042	0.023	0.061	0.063
	Q_5	0.03	0.029	0.021	0.291
	Q_6	0.117	0.029	0.084	0.000
	Q_7	-0.115	0.022	-0.171	0.000
	Q_8	0.032	0.024	0.025	0.179
	Q_9	0.343	0.020	0.305	0.000
	Q_{10}	0.183	0.028	0.126	0.000
	Q_{11}	0.300	0.012	0.404	0.000
Male	Intercept	-0.746	0.149		0.000
	Q_3	0.142	0.030	0.108	0.000
	Q_4	0.048	0.018	0.070	0.008
	Q_5	-0.015	0.027	-0.012	0.577
	Q_6	0.090	0.028	0.075	0.001
	Q_7	-0.087	0.017	-0.138	0.000
	Q_8	0.112	0.025	0.092	0.000
	Q_9	0.276	0.021	0.255	0.000
	Q_{10}	0.425	0.031	0.269	0.000
	Q_{11}	0.144	0.011	0.241	0.000

Finally, we modelled the teaching associates' SETs. The test of coincidence showed a significant effect, indicating that the regression line differed between the male and female teaching associates, F(8,7040) = 4.847, p <0.01. The second test, the test of intercepts, revealed that the intercepts in the regression did not significantly differ between genders, F(1,7040) = 29.964, p <0.01. Finally, the test of parallelism indicated that the slopes significantly differed, F(7,7040) = 5.522, p <0.01. The models for the female and male teaching associates are presented in Table 6.

Table 6

Regression models for the female and male teaching associates of the Faculty of Organizational Sciences

Gender of the teaching associate	Item	B	S.E.	β	Sig
Female	Intercept	0.913	0.083		0.000
	Q_3	0.180	0.019	0.188	0.000
	Q_5	0.080	0.015	0.102	0.000
	Q_6	0.246	0.020	0.256	0.000
	Q_8	0.054	0.014	0.071	0.000
	Q_9	0.073	0.016	0.080	0.000
	Q_{10}	0.156	0.019	0.147	0.000
	Q_{11}	0.028	0.006	0.065	0.000
Male	Intercept	0.407	0.049		0.000
	Q_3	0.218	0.013	0.222	0.000
	Q_5	0.103	0.011	0.125	0.000
	Q_6	0.289	0.013	0.305	0.000
	Q_8	0.030	0.010	0.039	0.000
	Q_9	0.091	0.012	0.096	0.000
	Q_{10}	0.170	0.012	0.173	0.000
	Q_{11}	0.025	0.004	0.056	0.000

The obtained R^2 is 0.467 for the female and 0.654 for the male teaching associates. Both models are statistically significant ($F_{female}(7,2699)=337.686$, $p<0.01$; $F_{male}(7,4341)=1174.090$, $p<0.01$). Although all the variables are statistically significant, the results should be taken with caution. Namely, Lin, Lucas and Shmueli (2013) observed that large samples could have a detrimental effect on the p-value, making it close to zero. Therefore, in linear regression modelling on large samples, all independent variables could be significant for the model, although, in fact, they might not be.

Limitations of the Study and Future Directions

There are some limitations of this study that should be pointed out. First, some factors influence the SET, which cannot be controlled and guaranteed. Namely, it is not possible to control the differences in the teaching styles of male and female lecturers and the impression they leave on their students. On the other hand, we cannot guarantee that these patterns would occur at another faculty or university or country, where student's expectations, attitudes, and socio-economic characteristics might be quite different. In other words, larger, multi-institutional samples would be valuable for a better understanding of patterns in the SET scores.

Additionally, the item of the unified SET limit the study in several ways. First, there is no information on the gender of the students, which could bring out more information on the issue of gender bias in the SET (Boring, 2015). The second limitation is that the item are highly subjective as they are related to the teaching style, way of

communicating with students, and grading standards. Therefore, the obtained results could be biased due to the halo effect which has been proved to occur by Shevlin et al. (2000). The third limitation of the questionnaire is the lack of information regarding the age of the student who fills in the SET. An interesting factor, whose impact could not be explored, is the ethnicity of the students and instructors (Dee, 2005).

Another limitation, which could also potentially distort the results, is the presence of the teacher/associate in the classroom while the students complete the SET. Namely, Feldman (1979) showed that the grades of instructors are higher if they are present in the classroom during the SET. Therefore, it is suggested that during the SET the lecturer should leave the classroom while the third person distributes and collects the forms (Centra, 2003). This suggestion could be implemented in the next cycle of the SET at the Faculty of Organizational Sciences.

On the other hand, there are several future directions that could be defined. One direction of research could be the alteration of the current SET used at the University of Belgrade. As mentioned above, the current SET does not include a question regarding the gender of the student and is mostly based on subjective items. The introduction of the gender-related question might provide additional information upon which more detailed analysis could be performed. Knowing the gender of the student, it could be explored whether it affects the scores given to male and female lecturers. The additional, more objective items could better depict the presence of gender bias towards lecturers.

Also, another direction of the study could be towards applying more advanced statistical methodologies, such as structural equation modelling (SEM) (Zhao & Gallant, 2012). For example, two SEM models per lecturer could be compared – one based on the results of female students, and the other based on the results of male students. Also, a similar approach could be used to create two SEM models per subject – one based on the results of female lecturers, and the other based on the results of male lecturers. Such approach could show which variables have different factor loadings, depending on the gender of the respondent/the gender of the lecturer.

The analysed SET is a unified SET conducted biannually at all 31 faculties of the University of Belgrade (UB, 2016). Therefore, it would be interesting to explore how the SET results vary among the faculties within the same field of science, whose subjects have similar curricula. Additionally, the analysis of the SET results through the years could be conducted at the Faculty or University level. Such an analysis has been carried out by Boring (2015) with success.

Discussion and Conclusion

The globalization of education initiated the reforms of higher education systems around the world, which have a goal to encourage flexibility in education systems. However, there is still need for a unified measurement of student satisfaction and teaching performance. The SET emerged as a much needed quantitative metric

(Simonacci & Gallo, 2017). The SET, on the one hand, makes the voice of students heard when it comes to the university affairs and, on the other hand, it provides the university administrators with an aura of accountability and legitimacy (Valsan & Sproule, 2008). Significant attention has been paid to the SET, its procedure, structure, and results, and the consequences of its results. Numerous studies related to the SET have been conducted in three major directions: on its validity (Zhao & Gallant, 2012), on its use by the faculty administration (Valsan & Sproule, 2008), and on the effects which can influence the final SET results (MacNell et al., 2015).

It is essential to discover the sources of bias in the SET and understand their impact on the SET result. One of the factors and sources of bias that has attracted significant attention of researchers is the lecturer's gender (Basow et al., 2006). Although the research results are inconclusive, understanding the potential gender bias in the SET scores is vital to the human resource management in academia. Namely, if the gender bias exists, the expectations of male and female lecturers differ significantly, so they are not to be evaluated using the same value system (Boring, 2015).

Herein, we set out to explore the results of the mandatory spring semester SET at the Faculty of Organizational Sciences, University of Belgrade, using the statistical multivariate analysis and data mining. Namely, data mining applied in the field of education is defined as educational data mining (EDM), which is slowly but surely being widely employed (Dobrota & Benković, 2014; Išljamović, Jeremić, & Lalić, 2016; Romero, Ventura, Pechenizky, & Baker, 2010). The employed Potthoff analysis has been used in several studies to compare the regression models between two or more groups (Lawson & Lips, 2014; Panaitescu et al., 2017). As such, it could be utilised to inspect whether there is gender bias in the SET scores. In other words, the idea was to create regression models based on the question scores as the independent variables and the overall lecturer score as the dependent variable, and to compare them between the male and female lecturers who teach the same subject.

Our approach was two-fold as we first attempted to analyse the presence of gender bias for six specific subjects and then on the level of the Faculty of Organizational Sciences. The comparison of models between genders showed that there was a difference between the male and female teachers of Mathematics 2 and Discrete Mathematical Structures, and between the male and female associates teaching Data Base. Our findings show interesting conclusions. When it comes to the regression models for female lecturers, the values of items related to measuring the "nurturing" traits were statistically significant, while the same items were not statistically significant in the models for male lecturers. It is also worth noting that, in the models for male lecturers, the values of items related to the measurement of unbiasness of grading and professionalism in the communication with students were statistically significant. In the continuation of our research, we compared the models on the level of the institution. The models for both teachers and teaching associates significantly differ. The models for teachers show notable results. If the female teachers are to improve

their SET scores, they need not improve the literature suggested and the amount of useful information they give to their students for future work, while their male colleagues should. This difference suggests that male teachers could embrace the "nurturing" traits. The difference in the models between the teaching associates is in the intensity of the impact of each item on the overall SET score.

This study aimed at providing additional insights into the issue of gender bias in the SET. More precisely, we attempted to explore how teaching style, communication skills, and grading differently impacted the overall SET score of the lecturer depending on the gender. This study suggests that, in some cases, the gender bias was detected, and that, based on the students' perceptions, the female teachers and associates should be more energetic, engaging, practical, respectful towards students, student-oriented, and supportive. On the other hand, the male lecturers should be objective and principled.

Herein, we demonstrated that different aspects of teaching are important for male and female teachers. Namely, students value different traits in male and female lecturers. Therefore, it is appropriate to note that the gender differentiated standards in the SET exist. The presented results could have multiple effects. First, they could draw the attention of university administration and national evaluation committees onto the issue of gender bias in the SET. If there is the difference, male and female teachers should be assessed differently, taking into account specific gender expectations. This information might act as a valuable contribution to the further development of SET and lecturer assessment. Second, the results could be a valuable feedback for lecturers as they can signal them which traits of their teaching could be enhanced. Also, the results could provide lecturers with more information on students' expectations and satisfaction with the teaching process. A combined effect is also possible as the results could raise the awareness of gender bias and try to break the gender stereotypes related to academia that exist among students, teaching staff, and administration.

We believe that our study might act as an impetus for further research on the validity of the SET, the effects of the gender bias, and on the SET currently conducted at the University of Belgrade.

Acknowledgement

This research is supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, as part of a project of technological development, project number ID III 47003.

References

Abel, M. H., & Meltzer, A. L. (2007). Student Ratings of a Male and Female Professors' Lecture on Sex Discrimination in the Workforce. *Sex Roles*, *57*(3–4), 173–180. https://doi.org/10.1007/s11199-007-9245-x

Aleamoni, L. M. (1999). Student Rating Myths Versus Research Facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*(2), 153–166. https://doi.org/10.1023/A:1008168421283

Arbuckle, J., & Williams, B. D. (2003). No Title. *Sex Roles*, *49*(9/10), 507–516. https://doi.org/10.1023/A:1025832707002

Attanasio, M., & Capursi, V. (Eds.)(2011). *Statistical Methods for the Evaluation of University Systems*. Heidelberg: Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-2375-2

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology*, *87*(4), 656–665. https://doi.org/10.1037//0022-0663.87.4.656

Basow, S. A. (2000). Gender dynamics in the classroom. In J. Chrisler, C. Golden, & P. D. Rozee (Eds.), *Lectures on the psychology of women* (pp. 44–55). New York: McGraw-Hill.

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. https://doi.org/10.1111/j.1471-6402.2006.00259.x

Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors: Evidence relating to the question of gender bias in teaching evaluation. *Journal of Educational Psychology*, *74*(2), 170–179. https://doi.org/10.1037//0022-0663.74.2.170

Boring, A. (2015). *Gender biases in student evaluations of teachers*. Documents de Travail de l'OFCE 2015-13, Observatoire Francais des Conjonctures Economiques (OFCE).

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

Centra, J. A. (2003). Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work? *Research in Higher Education*, *44*(5), 495–518. https://doi.org/10.1023/A:1025492407752

Chen, Yi., & Hoshower, L. B. (2003). Student Evaluation of Teaching Effectiveness: An assessment of student perception and motivation. *Assessment & Evaluation in Higher Education*, *28*(1), 71–88. https://doi.org/10.1080/02602930301683

Dee, T. S. (2005). A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? *American Economic Review*, *95*(2), 158–165. https://doi.org/10.1257/000282805774670446

Dobrota, M., & Benković, S. (2014). Comparing 'Ex catedra' and IT-Supported Teaching Methods and Techniques: Policy of Teaching Practice/Usporedba ex catedra i računalno potpomognutih metoda i tehnika poučavanja: politika nastavne prakse. *Croatian Journal of Education - Hrvatski časopis za odgoj i obrazovanje*, *16*(3), 91-108. https://doi.org/10.15516/cje.v16i0.581

Feldman, K. A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education*, *10*(2), 149–172. https://doi.org/10.1007/BF00976227

Ho, A., Watkins, D., & Kelly, M. (2001). The conceptual change approach to improving teaching and learning: An evaluation of a Hong Kong staff development programme. *Higher Education*, *42*(2), 143–169. https://doi.org/10.1023/A:1017546216800

Howell, D. (2013). *Statistical methods for psychology*. Belmont, CA: Cengage Wadsworth.

Išljamović, S. Z., Jeremić, V., & Lalić, S. (2016). Indicators of Study Success Related to Impact of University Students' Enrollment Status / Pokazatelji uspješnosti studiranja u korelaciji s tipom financiranja prilikom upisa na fakultet. *Croatian Journal of Education - Hrvatski časopis za odgoj i obrazovanje*, *18*(2), 583-606. https://doi.org/10.15516/cje.v18i2.1003

Jiang, Y. H., Javaad, S. S., & Golab, L. (2016). Data Mining of Undergraduate Course Evaluations. *Informatics in Education*, *15*(1), 85–102. https://doi.org/10.15388/infedu.2016.05

Johnson, R. (2000). The Authority of the Student Evaluation Questionnaire. *Teaching in Higher Education*, *5*(4), 419–434. https://doi.org/10.1080/713699176

Kogan, L. R., Schoenfeld-Tacher, R., & Hellyer, P. W. (2010). Student evaluations of teaching: perceptions of faculty based on gender, position, and rank. *Teaching in Higher Education*, *15*(6), 623–636. https://doi.org/10.1080/13562517.2010.491911

La Rocca, M., Parrella, M. L., Primerano, I., Sulis, I., & Vitale, M. P. (2017). An integrated strategy for the analysis of student evaluation of teaching: from descriptive measures to explanatory models. *Quality & Quantity*, *51*(2), 675–691. https://doi.org/10.1007/s11135-016-0432-0

Lawson, K. M., & Lips, H. M. (2014). The role of self-perceived agency and job attainability in women's impressions of successful women in masculine occupations. *Journal of Applied Social Psychology*, *44*(6), 433–441. https://doi.org/10.1111/jasp.12236

Lin, M., Lucas, H. C., & Shmueli, G. (2013). Research Commentary —Too Big to Fail: Large Samples and the p -Value Problem. *Information Systems Research*, *24*(4), 906–917. https://doi.org/10.1287/isre.2013.0480

Liu, O. L. (2012). Student Evaluation of Instruction: In the New Paradigm of Distance Education. *Research in Higher Education*, *53*(4), 471–486. https://doi.org/10.1007/s11162-011-9236-1

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), 291–303. https://doi.org/10.1007/s10755-014-9313-4

Maricic, M., Djokovic, A., & Jeremic, V. (2016). Gender bias in student assessment of teaching performance. In *Central European Conference on Information and Intelligent Systems* (pp. 137–143). Varaždin, Croatia.

Nowell, C., Gale, L. R., & Handley, B. (2010). Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. *Assessment & Evaluation in Higher Education*, *35*(4), 463–475. https://doi.org/10.1080/02602930902862875

Nunnally, B. H., & Bernstein, J. C. (1994). *Psychometric theory*. London, UK: McGraw-Hill.

OECD. (2009). *Teacher Evaluation: A Conceptual Framework and Examples of Country Practices*. Retrieved from http://www.oecd.org/education/school/44568106.pdf

Panaitescu, A. M., Akolekar, R., Kametas, N., Syngelaki, A., & Nicolaides, K. H. (2017). Impaired placentation in women with chronic hypertension who develop pre-eclampsia. *Ultrasound in Obstetrics & Gynecology*, *50*(4), 496–500. https://doi.org/10.1002/uog.17517

Potthoff, R. F. (1966). *Statistical aspects of the problem of biases in psychological tests*. Chapel Hill: University of North Carolina.

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile? *Quality Assurance in Education*, *15*(2), 178–191. https://doi.org/10.1108/09684880710748938

Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: the influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal*, *34*(1), 91–115. https://doi.org/10.1080/01411920701492043

Romero, C., Ventura, S., Pechenizky, M., & Baker, R. (2010). *Handbook of Educational Data Mining. Data Mining and Knowledge Discovery Series*. Boca Raton, Florida: CRC Press. https://doi.org/10.1201/b10274

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The Validity of Student Evaluation of Teaching in Higher Education: Love me, love my lectures? *Assessment & Evaluation in Higher Education*, *25*(4), 397–405. https://doi.org/10.1080/713611436

Simonacci, V., & Gallo, M. (2017). Statistical tools for student evaluation of academic educational quality. *Quality & Quantity*, *51*(2), 565–579. https://doi.org/10.1007/s11135-016-0425-z

Sinclair, L., & Kunda, Z. (2000). Motivated Stereotyping of Women: She's Fine if She Praised Me but Incompetent if She Criticized Me. *Personality and Social Psychology Bulletin*, *26*(11), 1329–1342. https://doi.org/10.1177/0146167200263002

Sprague, J., & Massoni, K. (2005). Student Evaluations and Gendered Expectations: What We Can't Count Can Hurt Us. *Sex Roles*, *53*(11–12), 779–793. https://doi.org/10.1007/s11199-005-8292-4

Truong, Y., Klink, R. R., Fort-Rioche, L., & Athaide, G. A. (2014). Consumer Response to Product Form in Technology-Based Industries. *Journal of Product Innovation Management*, *31*(4), 867–876. https://doi.org/10.1111/jpim.12128

UB. (2016). University of Belgrade. Retrieved from http://bg.ac.rs/en/index.php

Valsan, C., & Sproule, R. (2008). The Invisible Hands behind the Student Evaluation of Teaching: The Rise of the New Managerial Elite in the Governance of Higher Education. *Journal of Economic Issues*, *42*(4), 939–958. https://doi.org/10.1080/00213624.2008.11507197

Weaver, B., & Wuensch, K. L. (2013). SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients. *Behavior Research Methods*, *45*(3), 880–895. https://doi.org/10.3758/s13428-012-0289-7

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, *12*(1), 55–76. https://doi.org/10.1080/13562510601102131

Zhao, J., & Gallant, D. J. (2012). Student evaluation of instruction in higher education: exploring issues of validity and reliability. *Assessment & Evaluation in Higher Education*, *37*(2), 227–235. https://doi.org/10.1080/02602938.2010.523819

Milica Maričić
University of Belgrade, Faculty of Organizational Sciences,
Department of Operational Research and Statistics
Jove Ilića 154, 11000 Belgrade, Serbia
milica.maricic@fon.bg.ac.rs

Aleksandar Đoković
University of Belgrade, Faculty of Organizational Sciences,
Department of Operational Research and Statistics
Jove Ilića 154, 11000 Belgrade, Serbia
djokovic.aleksandar@fon.bg.ac.rs

Veljko Jeremić
University of Belgrade, Faculty of Organizational Sciences,
Department of Operational Research and Statistics
Jove Ilića 154, 11000 Belgrade, Serbia
veljko.jeremic@fon.bg.ac.rs

Appendix 1. Student Evaluation Questionnaire

Q_1. The classes are held regularly.
Q_2. The consultations are held regularly.
Q_3. Comprehensibility and presentation of the course syllabus.
Q_4. Compliance of the lecture and the scope of the subject.
Q_5. The teacher/associate encourages the student to take part in the class, to think critically, and to be creative.
Q_6. Lectures/classes assist students to master the course syllabus.
Q_7. Scope and quality of the suggested literature.
Q_8. The teacher/associate gives useful information for the students' future work.
Q_9. The teacher/associate answers students' questions and takes into account their comments.
Q_{10}. Professionalism in the communication with students.
Q_{11}. Objectivity and unbiasness in grading and evaluating the students' knowledge.
Q_{12}. The overall impression.

Valjanost studentskog vrednovanja nastave: postoji li spolna predrasuda?

Sažetak

Studentsko vrednovanje nastave (SVN) snažno, ali sigurno postaje važan alat za vrednovanje u visokom obrazovanju. Iako SVN daje povratnu informaciju o razini studentskog zadovoljstva nastavom i nastavnikom, vrijednost je takvih rezultata upitna. Nakon ekstenzivnih istraživanja vjeruje se da je spol nastavnika čimbenik koji iskrivljuje tako dobivene rezultate. U ovom je radu primijenjena Potthoff analiza kako bi se dodatno istražilo postoji li spolna predrasuda pri studentskom vrednovanju nastave. Ta se analiza, naime, vrlo uspješno koristila za usporedbu linearnih regresijskih modela između grupa. Cilj nam je bio izraditi modele sveukupnog dojma o nastavniku s pomoću nezavisnih varijabli koje se odnose na nastavu, komunikacijske vještine i ocjenjivanje, kao i usporediti ih ovisno o spolu. Dobiveni rezultati otkrivaju postojanje spolne predrasude u određenim slučajevima pri analiziranom studentskom vrednovanju nastave. Uvjereni smo da bi naše istraživanje moglo dati dodatni uvid u zanimljivu temu spolne predrasude pri studentskom vrednovanju nastave.

Ključne riječi: *Potthoff analiza; rudarenje edukacijskim podatcima; spolna predrasuda; studentsko vrednovanje nastave; visoko obrazovanje.*

Uvod

Od sedmog desetljeća prošlog stoljeća primjena studentskog vrednovanja nastave (SVN) dramatično je u porastu (Kogan, Schoenfeld-Tacher i Hellyer, 2010), pa je SVN, polako ali sigurno, postalo gotovo univerzalno prihvaćeno anketiranje s ciljem prikupljanja podataka od glavnih sveučilišnih dionika, studenata (Zabaleta, 2007). Danas su studentski upitnici za vrednovanje sasvim slični u cijelom svijetu ako se uzmu u obzir njihova struktura i pitanja. Većina ih se, naime, zasniva na Likertovoj ljestvici s pet ili sedam stupnjeva, pri čemu se od studenata zahtijeva da izraze svoje slaganje ili neslaganje s tvrdnjama o nastavnim materijalima, načinu poučavanja i predavaču (Kogan i sur., 2010). Nakon njihova manualnog ili računalnog prikupljanja, povratne informacije obično analiziraju nastavnici, odsjeci i fakulteti. Premda se taj proces na prvi pogled čini jednostavnim, u stvarnosti je daleko složeniji i podložan je

utjecaju unutarnjih i vanjskih čimbenika. Tri su glavna problema u vezi sa studentskim vrednovanjem nastave: uglavnom ordinalna vrsta podataka, subjektivnost pitanja i višestruka struktura studentskog zadovoljstva (Simonacci i Gallo, 2017).

Budući da omogućuje vrijedne podatke sveučilišnoj administraciji i vladinim tijelima, SVN postalo je obvezno u mnogim zemljama širom svijeta (OECD, 2009). Rezultati studentskog vrednovanja nastave koriste se, naime, u revizorskim i akreditacijskim postupcima unutar državnog i privatnog sektora (Johnson, 2000). Oni mogu također utjecati na napredovanje nastavnika u akademskom okruženju. U nekim su zemljama ti rezultati bitan sastavni dio akademskog životopisa te čine razliku među kandidatima pri zapošljavanju i napredovanju (Maricic, Djokovic i Jeremic, 2016).

Možemo primijetiti da rezultati studentskog vrednovanja nastave mogu višestruko utjecati na izvođenje sveučilišne nastave, akademsko osoblje i akreditacijski proces. SVN stoga predstavlja vrijedan izvor informacija o nastavi i zadovoljstvu studenata kolegijem ili nastavnikom (Chen i Hoshower, 2003). Imajući sve to u vidu, potrebno je pripremiti, provesti i tumačiti SVN oprezno i precizno. No, valjanost takvog vrednovanja dovode u pitanje mnogi stručnjaci u području vrednovanja nastave i nastavnika (Zhao i Gallant, 2012). Drugim riječima, nastavnici propituju kompetenciju studenata za vrednovanje nastavnika, nastave i silaba, a također su uvjereni da SVN ne raspolaže široko prihvaćenom definicijom koncepta „učinkovita nastava"(Johnson, 2000). S druge strane, stručnjaci su se prihvatili studentskog vrednovanja nastave na nacionalnoj i internacionalnoj razini (La Rocca, Parrella, Primerano, Sulis i Vitale, 2017). Glavni se problemi odnose na strukturu takvog vrednovanja i pitanja, kao i na valjanost studentskih odgovora koji su podložniji drugim čimbenicima nego nastavi (Maricic i sur., 2016; Zhao i Gallant, 2012).

Provedeno je ekstenzivno istraživanje kako bi se utvrdili čimbenici koji bi mogli imati učinak na studentsko ocjenjivanje nastavnika. Spol studenta, primjerice, utječe na rezultate. Basow (2000) ukazuje na to da spol studenta utječe na izbor „najboljeg" nastavnika i rezulate studentskog vrednovanja nastavnika. U jednom je drugom istraživanju Centra (2003) analizirao odnos između rezultata studentskog vrednovanja nastave i opterećenja, odnosno težine kolegija, dokazujući da su kolegiji koji su smatrani teškima uvijek slabije rangirani. Shevlin i njegovi suradnici (2000) također su pokazali da nastavnikova karizma lako utječe na rezultate studentskog vrednovanja nastave. Kao što je primjetno, čimbenici koji mogu iskriviti rezultate studentskog vrednovanja nastave pripadaju trima skupinama, ovisno o tome jesu li povezani sa studentom, kolegijem ili nastavnikom (Pounder, 2007).

Cilj je ovog istraživanja utvrditi utjecaj jednog čimbenika povezanog s nastavnikom na rezultate studentskog vrednovanja nastave, a to je nastavnikov spol. Stručnjaci u području visokog obrazovanja, sociologije i antropologije uvjereni su da je spol značajan čimbenik pri evaluaciji nastavnika, pa ga treba nastaviti istraživati (Basow, Phelan i Capotosto, 2006; MacNell, Driscoll i Hunt, 2015). Premda je literature o studentskom vrednovanju nastave znatno više posljednjih nekoliko desetljeća, nastojali

smo unaprijediti postojeća istraživanja. Glavna je pretpostavka u ovom istraživanju da različiti vidovi nastave imaju različit utjecaj na sveukupan rezultat nastavnika ovisno o njegovom spolu. Prvo, cilj nam je istražiti razlikuju li se nastavnici po spolu i predmetu kada je u pitanju utjecaj njihovih vještina i stila poučavanja na sveukupan rezultat studentskog vrednovanja nastave. Drugo, prikazujemo čimbenike koji utječu na sveukupan SVN rezultat nastavnika s obzirom na spol na razini određene visokoškolske institucije.

Rad počinje pregledom literature o spolu kao prepreci studentskom vrednovanju nastave. Potom se uvodi statistička metodologija koja je primijenjena da bi se istražila moguća spolna predrasuda. U središnjem dijelu rada opisano je provedeno studentsko vrednovanje nastave i prikazani su dobiveni rezultati. U posljednja dva dijela istaknuli smo potencijalna ograničenja našeg istraživanja, predložili smjernice za buduća istraživanja i detaljnije protumačili dobivene rezultate.

Spol kao predrasuda u studentskom vrednovanju nastave (SVN)

Da bi se bolje shvatila uočena pojava, donosimo kratak pregled literature o spolu kao predrasudi u studentskom vrednovanju nastave. Stavili smo žarište zanimanja na sljedeće teme: spol nastavnika, pozitivne i negativne osobine odabranih „najboljih" nastavnika, rezultati studentskog vrednovanja u odnosu na stil poučavanja, ocjene i sadržaj predavanja.

U istraživanjima čimbenika koji utječu na rezultate studentskog vrednovanja nastave spol je jedan od najzastupljenijih (Liu, 2012). Istraživanja na temu spolnih predrasuda iskristalizirala su se tijekom osamdesetih i devedesetih godina XX. stoljeća (Basow, 1995; Bennett, 1982). Tijekom godina ti su rezultati, međutim, ostali bez zaključka, a problem spolne predrasude neriješen. U nekim se istraživanjima spol spominje kao ograničavajući čimbenik (Basow, 1995; Maricic i sur., 2016), a u drugima se tvrdi kako takvog ograničenja nema te da su muški i ženski nastavnici vrednovani s pomoću sličnih parametara (Aleamoni, 1999; Arbuckle i Williams, 2003). Dakle, nepodudarni rezultati čine nejasnim odgovor na pitanje postoji li spolna predrasuda kada je riječ o rezultatima studentskog vrednovanja nastave, odnosno ima li spol nastavnika složenu i višestruku ulogu u rezultatima spomenutog vrednovanja (Basow i sur., 2006).

Sprague i Massoni (2005) proveli su zanimljivo istraživanje u kojem su zamolili studente i studentice da opišu svojeg „najboljeg" i „najgoreg" nastavnika. Kada se pogleda pet glavnih karakteristika „najboljeg" nastavnika, bez obzira na spol, rezultati pokazuju da je za nastavnike muškog spola važno tko *su* kao osobe, a da je za nastavnike ženskog spola bitno što *rade* tijekom nastave. To je važno otkriće jer pokazuje da su različiti nastavni pristupi poželjni za muške i ženske nastavnike, te da se isti napor različito vrednuje. Ukazali su, naime, na to da postoje značajne razlike u stilu poučavanja koje studenti vrednuju kod muških i ženskih nastavnika.

Jedan od čimbenika koji uvelike utječe na rezultate studentskog vrednovanja nastavnika jest stil poučavanja. Boring (2015) je u svojem istraživanju, naime, pokazala

da studenti različito pristupaju ocjenjivanju stila kojim poučavaju muški i ženski nastavnici. Njezino je opsežno istraživanje otkrilo da muški nastavnici dobivaju veće ocjene za one oblike poučavanja koji nisu vremenski zahtjevni, kao što su animacijske i prezentacijske vještine. No, situacija je potpuno drugačija u slučaju ženskih nastavnika koji su vrednovani za one oblike poučavanja koji zahtijevaju više vremena, kao što su priprema kolokvija, organizacija sata, povratna informacija i konzultacije.

Zanimljivo su istraživanje proveli Abel i Meltzer (2007) da bi utvrdili postoje li razlike u vrednovanju muškog i ženskog nastavnika nakon identičnog predavanja o spolno povezanoj temi. Pokazali su da su studenti vrednovali ženskog nastavnika i njezino predavanje kao više seksističko, te su dali sveukupno lošiju ocjenu iako je muški nastavnik održao isto predavanje. Dobiveni rezultat pokazuje da se spol nastavnika uzima u obzir pri sveukupnom ocjenjivanju predavanja i samog predavača. Međutim, ovo bi istraživanje moglo imati ograničenje jer je tema predavanja bila povezana sa spolom, što je moglo istaknuti spolne stereotipe među studentima.

Dvostruki se standardi primjenjuju kada se vrednuje objektivnost predavača pri ocjenjivanju studentskih postignuća. Ženske se nastavnike češće osuđuje, smatra nekompetentnim i kažnjava ih se lošijim SVN rezultatima u usporedbi s njihovim muškim kolegama kada studentima daju iste loše ocjene, tvrde Sinclair i Kunda (2000). Proveli su eksperiment u kojem su studenti morali vrednovati muške i ženske nastavnike nakon što su dobili rezultate testa. Rezultati su išli u prilog muškim nastavnicima/stručnjacima čak i kada su davali lošije ocjene i negativne povratne informacije. Drugim riječima, od ženskih se nastavnika očekuje da daju bolje ocjene, te pozitivnu povratnu informaciju i komentare (Boring, 2015).

Ako studenti različito ocjenjuju nastavnike na temelju spolnih stereotipa, tada muški i ženski nastavnici moraju različito poučavati, komunicirati i ocjenjivati studente da bi postigli bolje rezultate pri studentskom vrednovanju nastave. Sveučilišna administrativna tijela trebaju biti svjesna postojanja spolne predrasude u rezultatima studentskih vrednovanja nastave, te imati u vidu da se muški i ženski nastavnici nalaze pod opterećenjem kako bi odgovorili spolnim očekivanjima studenata (Sprague i Massoni, 2005). Prema prikazanoj literaturi, istraživanje je provedeno kako bi se utvrdilo utječe li spol bitno na vrednovanje nastave, komunikacije i cjelokupnog rezultata nastavnika.

Metodologija

Primjena multivarijatne analize i tehnike rudarenja podatcima pri ocjenjivanju SVN

Često se koristi linearno regresijsko modeliranje kako bi se otkrio odnos između specifičnih čestica u upitniku i sveukupnog nastavnog rezultata (Attanasio i Capursi, 2011). Naime, iako su čestice obično navedene u sklopu petostupanjske ili sedmostupanjske Likertove ljestvice, linearna se regresija primjenjuje s velikim uspjehom. Nowell, Gale i Handley (2010), primjerice, koristili su se raznim ordinalnim i skalnim varijablama da bi kreirali dva modela i objasnili prosjek pet karakteristika

nastavnika, te njegovo cjelovito vrednovanje. U novije su vrijeme Jiang i suradnici (2016) primijenili tehnike rudarenja podatcima i *multivarijatnu* regresiju da bi proveli longitudinalno istraživanje o preddiplomskom kolegiju na jednom velikom tehničkom fakultetu.

Osim linearne regresije koriste se i druge *multivarijatne* analize da bi se istražili rezultati studentskog vrednovanja nastave. Remedios i Lieberman (2008), primjerice, primijenili su analizu glavne komponentne (eng. *principal component analysis, PCA*) na analizu čimbenika koji utječu na rezultate studentskog vrednovanja nastave. Ho, Watkins i Kelly (2001) proveli su jednosmjernu multivarijatnu analizu varijance (MANOVA) da bi vrednovali učinke programa usavršavanja nastavnika na studentske pristupe učenju.

Dakle, možemo primijetiti da bi se različita multivarijatna analiza mogla primijeniti na rezultate studentskog vrednovanja nastave. No, linearna se regresija smatra dobrom metodom zbog nekoliko razloga. Prvo, kao što je navedeno, može se koristiti za evaluaciju i istraživanje studentskog vrednovanja nastave koje se uglavnom temelji na Likertovoj ljestvici s pet stupnjeva. SVN, ovdje pažljivo istraženo (detaljno objašnjeno u dijelu Provedeno istraživanje), ima takvu strukturu. Drugo, istraživačko pitanje nije postavljeno da bi se grupirale varijable i istražili čimbenici koji utječu na rezultate studentskog vrednovanja nastave, nego da bi se utvrdio i usporedio utjecaj određenih obilježja nastavnika. Konačno, postoji statistička analiza, Potthoff analiza, koja nam omogućuje vrlo uspješnu usporedbu regresijskih modela po grupama.

Potthoff analiza

Više knjiga i znanstveno-istraživačkih radova imalo je cilj uvesti metode i testove kako bi se usporedili koeficijenti iz običnih regresijskih modela najmanjih kvadrata (Howell, 2013; Potthoff, 1966). Potthoff analiza ističe se kao jednostavan način uspoređivanja linearnih regresijskih modela. Do sada je, naime, uspješno primjenjivana u socijalnoj psihologiji (Lawson i Lips, 2014), inovacijskom menadžmentu (Truong, Klink, Fort-Rioche i Athaide, 2014) i drugim područjima. Prednost te analize jest u činjenici da daje informaciju o tome (ne)razlikuju li se regresijski modeli koji su nastali na temelju različitih grupa i koristili se istim varijablama. Moguće je, naime, da je linearni regresijski model statistički važan za jednu grupu, ali ne i za neku drugu grupu. Osim toga, različiti koeficijenti mogli bi biti statistički značajni ovisno o grupi. Zamisao o uspoređivanju linearnih regresijskih modela, koja je još uvijek u fazi razvoja i dragocjena je za akademsku zajednicu, osobito Potthoff analiza, dokazuje nedavno objavljeni SAS i SPSS kod za povezane testove (Weaver i Wuensch, 2013).

Potthoff se analiza u osnovi sastoji od triju testova: podudarnosti, prekida i paralelizma. Test podudarnosti ima cilj utvrditi postoji li razlika između prekida, nagiba ili i jednog i drugog između promatranih grupa (Wuensch, 2016). Nadalje, test prekida određuje jesu li prekidi identični po grupama. Konačno, test paralelizma trebao bi dati informaciju o tome razlikuju li se nagibi značajno. Za više pojedinosti o analizi vidi u Wuensch (2016).

Nekoliko je prednosti Potthoff analize u istraživanju postojanja spolnog ograničenja. Prvo, odgovara na pitanje slabi li neka kategorijalna varijabla, u ovom slučaju spol nastavnika, odnos između neprekidnih varijabli (Lawson i Lips, 2014). Ta analiza, naime, daje uvid u to uzrokuju li nagibi ili prekidi ili oboje razlike između dviju regresijskih jednadžbi izvedenih za spol svakog nastavnika. Dakle, učinkovito uspoređuje regresijske modele. Drugo, razlika u regresijskim modelima i njihovim nagibima može biti dragocjen izvor informacija i za nastavnike i za sveučilišnu administraciju. Konačno, analiza se može provesti na ordinalnim podatcima (Lawson i Lips, 2014), koji se najčešće prikupljaju studentskim vrednovanjem nastave.

Provedeno istraživanje

SVN– Ispitanici, postupak, struktura

U istraživanju su sudjelovali studenti preddiplomskog studija na Fakultetu organizacijskih znanosti Sveučilišta u Beogradu. Svi su bili upisani kao redoviti studenti u jedan od četiriju dostupnih studijskih programa: Informacijski sustavi i tehnologije (IST), Menadžment, Operacijski menadžment i Upravljanje kvalitetom. Za potrebe naše analize koristili smo se rezultatima obveznog studentskog vrednovanja nastave u proljetnom semestru. Analizirano SVN je obvezno i unificirano anketiranje koje fakulteti u sastavu Sveučilišta u Beogradu provode dva puta godišnje na kraju svakog semestra. Studentima je dana mogućnost da vrednuju nastavnike i suradnike u nastavi čijim satima prisustvuju. SVN provedeno je za vrijeme predavanja u drugoj polovini svibnja 2016. godine. Volonter je podijelio ankete studentima u toku nastave, dok je nastavnik/suradnik bio u amfiteatru/učionici. Nakon toga su rezultati ankete uneseni s pomoću programa Blaise i statistički analizirani s pomoću programa SPSS 22. Konačno su rezultati studentskog vrednovanja nastave predočeni nastavniku/suradniku na kraju semestra, a pet je najboljih među njima (i nastavnika i suradnika) javno pohvaljeno na sjednici Fakultetskog vijeća.

SVN se sastojalo od triju dijelova. Prvi se dio odnosio na osnovne podatke o predmetu i nastavniku, kao što su: studijski program, vrednovani predmet, ime vrednovanog nastavnika i datum vrednovanja. U drugom je dijelu cilj bio prikupiti više podataka o studentima i njihovim ocjenama, pa su im postavljena pitanja o načinu financiranja studija (stipendija ili vlastiti izvor sredstava), dotadašnjem prosjeku ocjena, prethodnom slušanju kolegija, redovitom pohađanju nastave, tjednom broju sati namijenjenih učenju tog predmeta. Treći se dio razlikovao u pogledu vrednovanja nastavnika i suradnika. Sadržavao je 11 čestica (tvrdnji) za vrednovanje nastavnika i 9 čestica (tvrdnji) za vrednovanje suradnika za koje su studenti mogli izraziti svoje slaganje ili neslaganje na Likertovoj skali, pri čemu je 1 označavalo snažno neslaganje, a 5 snažno slaganje, dok je 0 bila u značenju bez odgovora. Pitanja koji su se koristila u ovom istraživanju ona su koja se odnose na vrednovanje nastavnika, a nalaze se u Prilogu 1. SVN primijenjeno za vrednovanje suradnika ne uključuje pitanja broj 4 (*Usklađenost nastavnika i raspona predmetnih tema)* i broj 7 (*Opseg i kvaliteta predložene literature*).

Rezultati

Cilj nam je u našoj analizi bio kreirati linearne regresijske modele za „*Ukupni dojam*" koristeći se svim pitanjima u vezi s vrednovanjem nastavnika. Naše je istraživanje bilo povezano s dvjema razinama: pojedinih sudionika i Fakulteta organizacijskih znanosti. Dobiveni su modeli poslije uspoređeni u odnosu na spol, pri čemu se koristila Potthoff analiza. Rezultati pokazuju postojanje razlika u različitim situacijama.

Od ukupno 66 predmeta čiji su predavači vrednovani, za našu smo analizu odabrali njih 6. U obzir nismo uzeli izborne kolegije jer bi vlastiti odabir studenata zakomplicirao analizu (Braga, Paccagnella i Pellizzari, 2014). Odabrani su kolegiji stoga morali biti obvezni ili za cijelu generaciju ili za cijeli studijski program. Osim toga, morali su obuhvaćati i muške i ženske nastavnike i/ili suradnike u nastavi, te imati više od 100 studentskih evaluacija. Potrebno je imati u vidu kako je jedan student mogao vrednovati više od jednog analiziranog predmeta i nastavnika ako je prisustvovao njihovoj nastavi. Ukupno smo analizirali 278 evaluacija za nastavnike (115 ženskih i 163 muška) i 1358 za suradnike u nastavi (648 ženske i 710 muške). Tablica 1 prikazuje odabrane predmete, vrstu vrednovanog nastavnika, broj studentskih vrednovanja i Cronbachov alpha za svakog sudionika po spolu. Izračunali smo Cronbachov alpha da bismo testirali konzistentnost studentskih odgovora. Zanimljivo je vidjeti kako većina analiziranih studentskih vrednovanja bilježi Cronbachov alpha ispod cutoff vrijednosti 0.7 (Nunnally i Bernstein, 1994), osim u slučaju Baze podataka za muške suradnike, Teorije odlučivanja za ženske suradnike, te Financijskog menadžmenta i računovodstva za ženske suradnike, što pak znači kako su ljestvice u većini slučajeva pouzdane, a odgovori konzistentni.

Tablica 1

Da bi se utvrdilo postoji li spolna predrasuda, primijenjena je Potthoff analiza. Prva se odnosi na analizu nastavnika za predmet Matematika 2. Test podudarnosti otkrio je značajan učinak, ukazujući na to da se regresijska linija razlikuje kod muških i ženskih nastavnika, $F(12,93) = 4,556$, $p < 0,01$, što je značilo da postoji razlika između tih dvaju regresijskih modela. Nastavili smo stoga analizirati da bismo utvrdili postoji li razlika u prekidu ili nagibima ili oboje. Drugi je test, test prekida, otkrio kako se prekidi u regresiji značajno ne razlikuju po spolu, $F(1,93) = 0,02$, $p > 0,05$. Konačno, test paralelizma otkrio je značajnu razliku kada su u pitanju nagibi, $F(11,93) = 4,935$, $p < 0,01$. Modeli za ženske i muške nastavnike prikazani su u Tablici 2. Dobiveni R^2 iznosi 0,809 za ženske nastavnike i 0,747 za muške nastavnike, a oba su modela statistički značajna ($F_{female}(11,38)=14,670, p<0,01$; $F_{male}(11,55)=14,759, p<0,01$). Ako se pažljivije pogledaju ti regresijski modeli (Tablica 2), vidi se da su različiti vidovi nastave važni ovisno o spolu. Za ženske nastavnike matematike važna su sljedeća obilježja: *redovito održavanje nastave* ($t=2,911$, $p<0,01$), *poticanje studenata na sudjelovanje u aktivnostima na satu, kritičko mišljenje i kreativnost* ($t=6,744$, $p<0,01$), *opseg i kvaliteta ponuđene literature* ($t=5,578$, $p<0,01$), *davanje korisnih informacija studentima za*

njihov budući rad (t=-2,837, p<0,01), *profesionalna komunikacija sa studentima* (t=4,344, p<0,01), te *objektivnost i nepristranost pri ocjenjivanju i vrednovanju znanja studenata* (t=-2,737, p<0,01). Taj rezultat pokazuje da su obilježja „odgajanja" visoko vrednovana u ženskih nastavnika, što je u skladu s rezultatom koji su dobili Sprague i Massoni (2005). No, kada je riječ o modelu za muške nastavnike matematike, onda su sljedeća obilježja značajna za ukupni dojam: *razumljivost i prezentacija silaba* (t=3,790, p<0,01), *profesionalna komunikacija sa studentima* (t=3,159, p<0,01), te *objektivnost i nepristranost pri ocjenjivanju i vrednovanju znanja studenata* (t=2,860, p<0,01).

Tablica 2

Sljedeće SVN koje smo modelirali bilo je ono za vrednovanje nastavnika koji su predavali kolegij Diskretne matematičke strukture. Razlikuje se od prethodnog modela jer se iz njega morala ukloniti varijabla *Nastava se redovito održava* s obzirom na to da su svi nastavnici imali isti rezultat. Analiza je pokazala da se ti modeli značajno razlikuju s obzirom na $F(11,139) = 117,264$, $p < 0,01$, te da postoji statistički značajna razlika i u prekidu ($F(1,139) = 81,356$, $p < 0,01$) i u nagibima ($F(10,139) = 108,954$, $p < 0,01$). Dva dobivena modela prikazana su u Tablici 3.

Tablica 3

Za ženske nastavnike R^2 iznosi 0,754 i 0,982 za muške nastavnike, dok su oba modela statistički značajna ($F_{female}(10,54)=16,587$, $p < 0,01$; $F_{male}(10,55)=458,960$, $p < 0,01$). U modelu za ženske nastavnike prekid je statistički značajan (t=14,941, p<0,01) skupa s varijablama *usklađenost predavanja i raspona predmetnih tema* (t=-3,122, p<0,01), *poticanje studenata na sudjelovanje u nastavi, kritičko mišljenje i kreativnost* (t=2,806, p<0,01), *pomoć studentu kroz nastavu pri svladavanju silabusa* (t=2,346, p<0,05), *davanje studentima korisnih informacija za budući rad* (t=4,371, p<0,01), te *profesionalna komunikacija sa studentima* (t=-2,660, p<0,01). Rezultati ponovno ukazuju na to da su obilježja „odgajanja" visoko vrednovana kada su ženski nastavnici u pitanju. U slučaju muških nastavnika matematike najmanje su dvije varijable statistički značajne (Q_9: t=2,126, p<0,05; Q_{10}: t=7,215, p<0,01), osim varijable *nastavnik odgovara na pitanja studenata i uvažava njihove komentare* (t=33,512, p<0,01).

Zatim smo razmotrili SVN za suradnike u nastavi kolegija Uvod u informacijske sustave. Test podudarnosti nije pokazao razliku između tih dvaju modela ($F(10,93)=1,293$, $p > 0,05$). Budući da se međusobno ne razlikuju, nisu provedena preostala dva testa, prekida i paralelizma.

Potom je modeliran Ukupni dojam za suradnike u nastavi kolegija Baze podataka. Utvrđena je pritom značajna razlika između dvaju modela, ($F(10,377)=4,080$, $p < 0,01$), i to kada je riječ o nagibima ($F(9,377)=4,452$, $p < 0,01$), a ne o prekidu ($F(1,377)=0,032$, $p > 0,05$). Dobiveni R^2 iznosi 0,602 za ženske suradnike i 0,460 za muške suradnike, a oba su modela statistički značajna ($F_{female}(9,201)=33,834$, $p < 0,01$; $F_{male}(9,176)=16,600$, $p < 0,01$). Oba su modela prikazana u Tablici 4. Ako se bolje pogleda model za ženske

suradnike, onda se može uočiti kako su, osim prekida (t=-2,153, p<0,05), značajne tri varijable: *sposobnost poticanja studenata na aktivno sudjelovanje u nastavi, kritičko mišljenje i kreativnost* (t=2,579, p<0,01), *profesionalnost u komunikaciji sa studentima* (t=6,864, p<0,01), te *objektivnost i nepristranost u ocjenjivanju i vrednovanju znanja studenata* (t=2,730, p<0,01). Model za muške suradnike ukazuje na statistički značaj sljedećih varijabli: *razumljivost i prezentacija silaba* (t=2,382, p<0,05), *pomoć studentima kroz nastavu da svladaju silab* (t=3,119, p<0,01), te *objektivnost i nepristranost u ocjenjivanju i vrednovanju znanja studenata* (t=2,246, p<0,05).

Tablica 4

Nadalje, ukupan dojam u slučaju suradnika u nastavi kolegija Teorija odlučivanja modeliran je kao model za nastavnike kolegija Diskretne matematičke strukture jer je varijablu *Nastava se redovito održava* ponovno trebalo ukloniti. Između dvaju modela nije utvrđena nikakva značajna razlika, F(9,340)=1,874, p>0,05). Konačno, vrednovali smo SVN-ove za suradnike u nastavi kolegija Financijski menadžment i računovodstvo. Zanimljivo je kako nije utvrđena statistički značajna razlika između dvaju modela (F(10,169)=1,396, p >0,05).

Da bismo dodatno istražili postoji li spolna predrasuda kada je u pitanju SVN provedeno

na Fakultetu organizacijskih znanosti, proveli smo istu analizu za nastavnike i suradnike na razini institucije. Analizirali smo studentsko vrednovanje 4162 nastavnika i 7056 suradnika. No, u ovoj su analizi izostavljena pitanja *Nastava se redovito održava* i *Konzultacije se redovito održavaju* zbog visoke frekvencije rezultata 5.

Prvo, izrađen je model za ukupan nastavnički dojam. Potthoff analizom utvrđeno je kako postoji statistički značajna razlika između dvaju modela kao F(10,4142) = 14,314, p <0,01, da je razlika u prekidu (F(1,4142) = 5,447, p <0,01), a ne u nagibu (F(9,4142) = 14,869, p <0,01). Modeli za ženske i muške nastavnike prikazani su u Tablici 5. Dobiveni R^2 iznosi 0,517 za ženske i 0,477 za muške nastavnike. Oba su modela statistički značajna (F_{female}(9,2211)=263,006, p <0,01; F_{male}(9,1931)=195,867, p <0,01). Zanimljivo je postojanje šest obilježja koja su važna za ženske i muške nastavnike: Q3 (t_{female}=3,151, p <0,01; t_{male}=4,772, p <0,01), Q6 (t_{female}=4,066, p <0,01; t_{male}=3,192, p <0,01), Q7 (t_{female}=-5,260, p <0,01; t_{male}=-5,184, p <0,01), Q9 (t_{female}=17,053, p <0,01; t_{male}=12,888, p <0,01), Q10 (t_{female}=6,454, p <0,01; t_{male}=13,600, p <0,01), te Q11 (t_{female}=24,578, p <0,01; t_{male}=13,363, p <0,01). Za muške nastavnike, osim tih šest obilježja, statistički su važni *usklađenost predavanja i raspona predmetnih tema* (t=2,671, p<0,01) i *davanje studentima korisnih informacija za njihov budući rad* (t=4,419, p<0,01). Još jedan rezultat privlači pozornost, a to je kako kvaliteta preporučene literature značajno smanjuje sveukupan rezultat studentskog vrednovanja nastave i za muške i za ženske nastavnike.

Tablica 5

Konačno smo modelirali studentsko vrednovanje nastave za suradnike u nastavi. Test podudarnosti pokazao je značajan učinak, ukazujući na to da se regresijska linija razlikuje između muških i ženskih suradnika, $F(8,7040) = 4,847$, $p < 0,01$. Drugi test, test prekida, otkrio je kako se prekidi u regresiji značajno ne razlikuju u odnosu na spol, $F(1,7040) = 29,964$, $p < 0,01$. Na kraju je test paralelizma pokazao da se nagibi značajno razlikuju, $F(7,7040) = 5,522$, $p < 0,01$. Modeli za ženske i muške suradnike prikazani su u Tablici 6.

Tablica 6

Dobiveni R^2 iznosi 0,467 za ženske i 0,654 za muške suradnike u nastavi. Oba su modela statistički značajna ($F_{female}(7,2699)=337,686$, $p < 0,01$; $F_{male}(7,4341)=1174,090$, $p < 0,01$). Iako su sve varijable statistički značajne, rezultate bi trebalo uzeti u obzir oprezno. Lin, Lucas i Shmueli (2013) su, naime, primijetili da veliki uzorci mogu imati detrimentalni učinak na p-vrijednost, približavajući je nuli. U slučaju izrade linearnih regresijskih modela na velikim uzorcima, sve bi nezavisne varijable stoga trebale biti značajne za model iako to zapravo možda i nisu.

Ograničenja i buduće smjernice za istraživanje

Postoje određena ograničenja u ovom istraživanju koja treba istaknuti. Prvo, određeni čimbenici utječu na SVN, što se ne može ni kontrolirati ni jamčiti. Nije moguće, naime, kontrolirati razlike u stilu poučavanja muških i ženskih nastavnika te dojam koji ostavljaju na studente. S druge strane, ne možemo jamčiti da bi se slični obrasci pojavili na nekom drugom fakultetu ili sveučilištu ili državi gdje bi studentska očekivanja, stavovi i socioekonomski uvjeti mogli biti sasvim drugačiji. Veći, višeinstitucionalni uzorci bili bi, naime, vredniji za bolje razumijevanje obrazaca pronađenih u rezultatima studentskog vrednovanja nastave.

Osim toga, pitanja unificiranog studentskog vrednovanja nastave ograničavaju istraživanje na nekoliko načina. Prvo, nema podataka o spolu studenata, što bi moglo pridonijeti razmatranju spolne predrasude pri studentskom vrednovanju nastave (Boring, 2015). Drugo, pitanja su izrazito subjektivna jer se odnose na stil poučavanja, način komunikacije sa studentima i standarde ocjenjivanja. Dakle, dobiveni bi rezultati mogli biti neobjektivni zbog halo učinka, što se pokazalo u istraživanju Shevlin i suradnika (2000). Treće, u upitniku nedostaje podatak o dobi studenta koji ga ispunjava. Zanimljivi čimbenik, čiji se učinak nije mogao istražiti, jest etnička pripadnost studenata i predavača (Dee, 2005).

Dodatno ograničenje koje bi moglo utjecati na rezultate odnosi se na nazočnost nastavnika/suradnika u učionici tijekom ispunjavanja anketnih listića. Feldman (1979) je pokazao, naime, kako su predavači ocjenjivani višim ocjenama ako su bili u učionici dok su studenti odgovarali na pitanja iz upitnika. Stoga se preporučuje da nastavnik napusti učionicu dok treća osoba dijeli i prikuplja anketne listiće (Centra, 2003). Ta bi se preporuka trebala uvažiti u sljedećem ciklusu studentskog vrednovanja nastave na Fakultetu organizacijskih znanosti.

Međutim, moglo bi se utvrditi nekoliko smjernica za buduća istraživanja. Jedan pravac istraživanja mogao bi se odnositi na promjenu u sadašnjem studentskom vrednovanju nastave na Sveučilištu u Beogradu. Kao što je već navedeno, ono danas ne sadrži pitanje o spolnoj pripadnosti studenta i uglavnom se temelji na subjektivnim pitanjima. Uvođenjem spolno referentnog pitanja mogla bi se dati dodatna informacija, što bi omogućilo detaljniju analizu. Znajući spol studenata, moglo bi se istražiti utječe li njihov spol na rezultate koji vrijede za muške i ženske nastavnike. Dodatna bi objektivnija pitanja mogla dati bolju sliku o prisutnosti spolne predrasude o nastavnicima.

Nadalje, u drugom bi se pravcu istraživanja mogla zasnivati na uporabi naprednijih statističkih metodologija, kao što je modeliranje strukturnih jednadžbi (eng. *structural equation modelling*, SEM) (Zhao i Gallant, 2012). Primjerice, mogla bi se uspoređivati dva SEM modela po nastavniku – jedan utemeljen na rezultatima ženskih studenata, a drugi onih muških. Sličan bi se pristup također mogao primijeniti na izradu dvaju SEM modela po predmetu – jedan zasnovan na rezultatima ženskih nastavnika, a drugi muških. Takav bi pristup mogao pokazati koje varijable imaju različita faktorska opterećenja ovisno o spolu studenta/spolu nastavnika.

Analizirano SVN unificirano je studentsko vrednovanje nastave koje se provodi svake druge godine na svakom od 31 fakulteta Sveučilišta u Beogradu (UB, 2016). Bilo bi, dakle, zanimljivo istražiti kako se rezultati studentskog vrednovanja nastave razlikuju po fakultetima unutar istog znanstvenog područja čiji predmeti imaju slične kurikule. Osim toga, analiza rezultata studentskog vrednovanja nastave tijekom godina mogla bi se provesti na fakultetskoj ili sveučilišnoj razini. Takvu je analizu uspješno proveo Boring (2015).

Rasprava i zaključak

Globalizacija obrazovanja unijela je reforme u sustave visokog školstva širom svijeta s ciljem poticanja njihove fleksibilnosti. No, još uvijek postoji potreba za unificiranim mjerenjem studentskog zadovoljstva i izvođenja nastave. SVN pojavilo se kao itekako nužno kvantitativno metričko rješenje (Simonacci i Gallo, 2017). Međutim, SVN omogućuje da se čuje glas studenata o sveučilišnim aktivnostima, ali ujedno daje onima koji vode sveučilište auru odgovornosti i legitimiteta (Valsan i Sproule, 2008). Znatna je pozornost usmjerena na SVN, njegovo provođenje, strukturu, rezultate i posljedicu tih rezultata. Provedena su brojna istraživanja u vezi sa studentskim vrednovanjem nastave i to u tri glavna pravca: njegova validnost (Zhao i Gallant, 2012), njegovo korištenje od fakultetske uprave (Valsan i Sproule, 2008), te učinci koji mogu utjecati na krajnje rezultate studentskog vrednovanja nastave (MacNell i sur., 2015).

Ključno je otkriti izvore predrasuda pri studentskom vrednovanju nastave i shvatiti njihove utjecaje na rezultate studentskog vrednovanja nastave. Jedan od takvih čimbenika i izvora koji privlači znatnu pozornost istraživača predstavlja spol predavača (Basow i sur., 2006). Iako rezultati istraživanja nisu sveobuhvatni, razumijevanje

potencijalne spolne predrasude u rezultatima studentskog vrednovanja nastave vitalno je kada je u pitanju upravljanje ljudskim resursima u akademskoj zajednici. Naime, ako postoji spolno ograničenje, onda se značajno razlikuju očekivanja muških i ženskih nastavnika, pa se ne mogu ni vrednovati istim sustavom vrijednosti (Boring, 2015).

Ovdje smo istražili rezultate obveznog studentskog vrednovanja nastave u proljetnom semestru na Fakultetu organizacijskih znanosti Sveučilišta u Beogradu koristeći se statističkom multivarijatnom analizom i rudarenjem podatcima. Naime, rudarenje podatcima primijenjeno u području obrazovanja definira se kao rudarenje edukacijskim podatcima (eng. *educational data mining*, EDM), a polako se iako sigurno sve više koristi (Dobrota i Benković, 2014; Išljamović, Jeremić i Lalić, 2016; Romero, Ventura, Pechenizky i Baker, 2010). Primijenjena Potthoff analiza koristila se u nekoliko istraživanja za usporedbu regresijskih modela između dviju ili više grupa (Lawson i Lips, 2014; Panaitescu i sur., 2017). Kao takva mogla bi se koristiti za utvrđivanje (ne) postojanja spolnog ograničenja kada su u pitanju rezultati studentskog vrednovanja nastave. Namjera je bila zapravo kreirati regresijske modele zasnovane na rezultatima proizašlim iz odgovora na pitanja kao nezavisnim varijablama i sveukupnom rezultatu nastavnika kao zavisnoj varijabli, te usporediti ih između ženskih i muških predavača koji poučavaju isti predmet.

Naš je pristup dvostruk jer smo najprije nastojali analizirati prisutnost spolne predrasude za šest specifičnih predmeta, a onda na razini Fakulteta organizacijskih znanosti. Spolna usporedba modela pokazala je da postoji razlika između muških i ženskih nastavnika koji poučavaju Matematiku 2 i Diskretne matematičke strukture, te između muških i ženskih suradnika u nastavi kolegija Baze podataka. Naši rezultati dovode do zanimljivih zaključaka. Kad je riječ o regresijskim modelima za ženske nastavnike, statistički su značajne vrijednosti za pitanja koja se povezuju s mjerenjem karakteristika „odgajanja", a ista pitanja nisu statistički značajna u modelima za muške nastavnike. Bitno je također primijetiti da su u modelima za muške nastavnike vrijednosti za pitanja kojima se mjere nepristranost pri ocjenjivanju i profesionalnost komunikacije sa studentima statistički značajne. U nastavku našeg istraživanja usporedili smo modele na razini institucije. Modeli nastavnika i suradnika u nastavi statistički se značajno razlikuju. Modeli za nastavnike pokazuju istaknute rezultate. Ako nastavnice trebaju poboljšati svoj rezultat postignut pri studentskom vrednovanju nastave, trebaju poboljšati predloženu literaturu i količinu korisnih informacija koje daju studentima za budući rad, a to ne trebaju činiti njihovi muške kolege. Spomenuta razlika sugerira da bi muški nastavnici trebali prigrliti obilježja „odgajanja". Razlika u modelima između suradnika u nastavi nalazi se u jačini utjecaja svakog pitanja na sveukupan rezultat studentskog vrednovanja nastave.

Ovo je istraživanje imalo cilj dati dodatni uvid u studentsko vrednovanje nastave s obzirom na spolnu predrasudu. Točnije, nastojali smo istražiti kako stil poučavanja, komunikacijske vještine i ocjenjivanje različito utječu na sveukupan rezultat nastavnika pri studentskom vrednovanju nastave ovisno o spolu. Ovo istraživanje sugerira

postojanje spolnog ograničenja u nekim situacijama, te da – na temelju studentske percepcije – ženski nastavnici i suradnici u nastavi trebaju biti energičniji, angažiraniji, praktičniji, više poštovati studente, više im biti usmjereni i više ih podupirati. S druge strane, muški bi nastavnici trebali biti objektivniji i principjelniji.

Stoga je ovdje ukazano na to kako su različiti aspekti poučavanja važni za muške i ženske nastavnike. Studenti, naime, vrednuju različita obilježja muških i ženskih nastavnika. Odgovarajuće je stoga primijetiti da postoje spolno različiti standardi kada je u pitanju SVN. Prikazani bi rezultati mogli imati višestruke učinke. Prvo, mogli bi privući pozornost sveučilišne uprave i nacionalnih evaluacijskih povjerenstava na problem spolnog ograničenja pri SVN. Ako razlika postoji, onda bi muške i ženske nastavnike trebalo vrednovati različito imajući u vidu specifična spolna očekivanja. Ta bi informacija mogla djelovati kao vrijedan doprinos daljnjem razvoju studentskog vrednovanja nastave i nastavnika. Drugo, taj bi rezultat mogao biti vrijedna povratna informacija nastavnicima jer bi im eventualno signalizirao koje bi karakteristike svojeg poučavanja mogli unaprijediti. Osim toga, prikazani bi rezultati mogli nastavnicima dati više informacija o studentskim očekivanjima i zadovoljstvu nastavnim procesom. Moguć je također kombinirani učinak jer bi nas rezultati osvijestili o postojanju spolnog ograničenja, te bi se njima nastojalo dokinuti spolne stereotipe u akademskom svijetu među studentima, nastavnicima i upravom.

Uvjereni smo da bi naše istraživanje moglo dati poticaj daljnjim istraživanjima valjanosti studentskog vrednovanja nastave, učinaka spolnog ograničenja i SVN kakvo se trenutno provodi na Sveučilištu u Beogradu.

Zahvala

Ovo je istraživanje provedeno uz potporu Ministarstva obrazovanja, znanosti i tehnološkog razvoja Republike Srbije kao dio projekta tehnološkog razvoja broj projekta ID III 47003.