Gender bias in student assessment of teaching performance

Milica Maricic, Aleksandar Djokovic, Veljko Jeremic

Faculty of Organizational Sciences

University of Belgrade

Jove Ilića 154, Belgrade

{milica.maricic, djokovic.aleksandar, jeremic.veljko}@fon.bg.ac.rs

Abstract. *Student evaluation of teaching (SET) is a worldwide, commonly used measure of teaching performance. Namely, higher educational institutions are in some countries, such as in Serbia, obliged to perform biannual SETs. Although the procedures and instruments of data collection vary, the purpose remains the same: get feedback on the quality of teaching from the universities' important stakeholder, the students themselves. The obtained results are usually purely analysed with the goal of further improvement of both the teaching process and the teacher's performance. However, an interesting question emerges: are students unbiased when evaluating their teachers? This study attempts to address the issue of gender bias in annual SET at the University of Belgrade, Faculty of Organizational Sciences. The presented analysis aimed to investigate whether there is difference in the SET scores between male and female teachers. We believe this study might shed light on how students are (un)biased when doing SET regarding the gender effect.*

Keywords. Student evaluation of teaching, Higher education, Gender bias, University of Belgrade

1 Introduction

Since the 1970's the use of student evaluation of teaching (SET) has expanded dramatically (Kogan, Schoenfeld-Tacher & Hellyer, 2010). Nowadays, evaluations of teachers and courses are performed at the end of semesters in most higher education institutions (Subramanya, 2014). Traditional SETs are usually comprised of a series of open and closed questions about course content, suggested literature, teaching effectiveness, and learning experience. Closed questions are generally created using the five- or seven-point Likert scale. Most commonly, the surveys are anonymous and are hand distributed while the evaluated teacher might or might not be present. The SETs and the process of conducting SET mostly remained unchanged since they first appeared.

Although SETs were originally intended for informative purposes, such surveys came into use for faculty personnel decisions (Galbraith, Merrill & Kline, 2012). Administrators routinely consider SET scores in hiring, tenure, promotion, and salary decisions (Laube et al., 2007). At some higher educational institutions, teachers' results of the annual SET are perceived as relevant additional qualifications besides research and mentoring. Also, SET results are used in both national or private auditing procedures of the educational institution (Johnson, 2000). As it can be noted, the SET results are considered a valuable source of information upon which important decisions are made. Namely, the results of SETs can severely influence a teacher's career, especially young teaching associate's career who has recently become a lecturer, and the institution's quality certificates.

Having the above-mentioned in mind, we can conclude that SETs should be carefully conducted and designed to preserve their validity. Namely, potential bias in SET scores is a matter of great importance (MacNeel et al., 2015). Marsh and Roche (1997) state that SETs are a valid multidimensional measure whose results could be used in decision making. However, SETs are extremely difficult to validate since there is no single criterion of teaching effectiveness and satisfaction (Marsh, 2007). Also, there is no consensus on the attitude of teachers towards the validity of SETs (Nasser & Fresko, 2010). Teachers' main doubts are whether students are mature enough to evaluate the quality of teaching, whether students can evaluate the course right after it has finished or a certain amount of time has to pass, and the influence of other factors on scores such as difficulty of the course, grading, teacher popularity (Aleamoni, 1987; Spiller & Ferguson, 2011).

Following the teachers' concerns, the academic community has risen serious questions regarding the validity of SETs. First, it is still unclear what are the dimensions of teaching effectiveness that the SET should measure. For example, Swartz et al. (1999) identified two factors, while on the other hand, Marsh and Dunkin (1992) suggested nine factors. Secondly, several authors raised doubts on the current

methodology of the evaluation tests (Hampton & Reiser, 2004; Marsh, 2007). Namely, the form and questions on SETs have not significantly changed throughout the years (Subramanya, 2014). On the other hand, teaching and learning techniques have undergone tremendous modifications, especially due to the increased ICT use (Bingimlas, 2009). Thus, the currently employed SETs might not be an accurate and objective measurement. Finally, students are susceptible to other factors rather than teaching when grading teachers. For example, in their study, Shevlin and associates (2000) showed that student evaluations are easily influenced by charisma of the teacher. Also, Clayson (2013) revealed that first impressions affect the final evaluations of students. In a thorough literature analysis, Pounder (2007) divides the factors which can distort the SET results into three groups: student related, course related, and teacher related factors. Therefore, we can conclude that the results of SETs might not be sufficiently reliable and valid indicators of teacher effectiveness (Baker et al., 2010).

As it can be noted, there are various factors which have a direct or indirect effect on the SET scores. The aim of this study is to explore the influence of the teacher's gender on the SET scores. The primary question is whether students perceive any differences between male and female teachers who teach the same subject. The gender schema might impact student's perceptions and expectations of their teachers. The academia has been puzzled by this question for several decades. On one hand, several studies show that there are differences between the grades assigned to male and female teachers (Bachen et al., 1999; Sprague & Massoni, 2005). On the other hand, Feldman (1993) reported no or little gender bias in his research, while Aleamoni (1999) also found no relationship between either teacher gender or student gender and student SET scores. Sociologist specialised in gender studies highly disagree that gender is not a significant factor in teacher evaluations (Laube et al., 2007). Therefore, it can be observed that it is still unclear whether gender bias exists in SET scores (Basow, 2000a) and whether gender plays a complex and multifaceted role in SET scores (Basow et al., 2006).

2 Gender bias in student evaluation of teaching (SET)

To better understand the effects of gender bias herein we provide an overview of the recent literature. The topics we place our attention on are the gender of the student, the gender of the lecturer, the positive and negative traits of chosen teachers, and the SET scores.

Interesting studies were conducted by Basow (2000b) and Sprague and Massoni (2005). Namely, both studies asked students to depict their "best" and "worst" male and female teachers. Their results were conclusive. Traits of "best" female teachers were

caring and nurturing, while the traits of "best" male teachers were funny and entertaining. On the other hand, when it comes to "worst" teachers, common traits for both genders were unorganised, unclear, indifferent, and rude. "Worst" female teachers were just the opposite of the "best" female teacher; they were characterised as rigid, mean, and unfair. Interestingly, "worst" male teachers were self-centred and unenthusiastic. Also, Kierstead, D'Agostino, and Dill (1988) found that women teachers received higher scores if they were friendly than if they were unfriendly, whereas this characteristic did not affect the scores of men teachers. Women and men teachers, therefore, have to meet different expectations of students which are related to their gender (Sprague & Massoni, 2005). Women teachers are especially affected by the observed stereotypes. Students seem to expect more nurturing behaviour from them, but they, in turn, often judge that behaviour to be less professorial (Anderson & Smith, 2005). Contrarily, if women teachers fail to meet students' expectations of women, they are characterised as too masculine and rigorous (Valian, 1998).

It is also of interest to analyse how the gender of the respondent affects the gender of the "best" teacher and the SET scores. When asked to name their best teacher male students did not choose female teachers as often as female students did (Basow, 2000b). Bachen et al. (1999) initiated research on the interaction between students' and teachers' gender in regard to SET scores. They showed that female students gave higher grades to their female teachers. Contrarily, their male colleagues' evaluations of male and female teachers did not differ significantly. One more interesting study regarding the assigned scores of male and female teachers should be mentioned. Namely, MacNell et al. (2015) showed that regardless of performance, students rated the female instructors significantly more rigorously than the male instructors, which suggests that a female instructor would have to work harder than a male to receive comparable ratings. On the other hand, in the study by Cashin (1999) students rated women higher than men.

The results of the presented studies provide evidence that there might be some bias against female teachers and that students have gendered expectations. The administration should be aware of the gender bias when analysing SET scores and should have in mind that both men and women teachers are under burden to answer to students' gendered expectations (Sprague & Massoni, 2005).

3. Conducted research

The following section sees the description of the SET distributed at the University of Belgrade, Faculty of Organizational Sciences, its participants, and the procedure of the SET distribution. Also, the phases

of the research which had the aim to explore whether there is gender bias in students' assessment of teacher performance are presented. Finally, the obtained results are given.

3.1 Procedure, participants, and the SET

Participants of the conducted study were undergraduate students of the University of Belgrade's Faculty of Organizational Sciences (FOS). They were all enrolled on full-time courses on one of four study programmes available: Information Systems and Technologies (IS&T), Management, Operational Management, and Quality Management. Herein, for our analysis, we used the results of the summer midterm SET. The conducted SET is a mandatory and unified survey all faculties of the University of Belgrade administer twice a year. FOS students were given an opportunity to rate the teachers and teaching associates whose classes they have attended. The SET was distributed on classes in the second half of May 2016. They were administered to students by a volunteer during class while the instructor was in the 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 lecturer after the end of the semester while the best five teachers and five teaching associates were commended on the faculty council.

First four questions of the SET regard the study programme, the subject the evaluated teacher teaches, the name of the teacher and the date of the evaluation. The next six questions aim at depicting the students' achievements so far: whether they are self-financed or on a state scholarship, their average grade, have they attended the course before, have they regularly attended the classes and the number of hours they weekly spent doing the subject. The following set of questions is crucial for the SET as it is related to the evaluation of the teacher or the associate. Namely, the questionnaires used to assess teachers and associates are different. It consists of 11 statements when assessing teachers and 9 statements when assessing associates on which the students should express their agreement or disagreement on a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree) including 0 (No answer).

Most of the questions on the SET regard the lecturer's technique of teaching and grading: whether he/she is biased when grading, is he/she professional when communicating with students, whether he/she is giving useful information for further research, and whether his/her classes motivate students for the future studies. Namely, the level of agreement with this statements solely depends on the specific way the lecturer organises and conducts his classes. As herein we aim to assess the impact of the lecturers' gender on the grades he/she received we decided to place our attention on the few questions which are not related to

the teaching technique. When assessing teachers, we placed our attention on three questions: "Compliance of the lecture and the scope of the subject", "Scope and quality of the suggested literature", and "Overall impression". On the other hand, the associate's SET does not have the two questions regarding the compliance of the lecture and the suggested literature. Therefore, when analysing their results, we only explored the answers to the question "Overall impression".

3.2 Results

Out of 66 subjects whose lectures were assessed for our analysis, we chose nine subjects. The selected subjects had to have both male and female teachers and/or associates and more than 100 students' assessments. Table 1 presents the chosen subjects, the type of assessed lecturer and the number of student assessments. One must have in mind that one student could have assessed more than one of the observed subjects and lecturers if he had attended their classes. In total, we observed 785 teachers' assessments and 2368 associates' assessments.

Table 1: The chosen subjects, type of lecturer assessed, and the number of assessments

Subject	Lecturer	Assessments
Business processes modelling	Associate	311
Data base	Associate	397
Decision theory	Associate	358
Discrete mathematical structures	Teacher	161
Financial management and accounting	Associate	189
Introduction to information systems	Teacher	213
Discrete mathematical structures	Associate	414
Mathematics 2	Teacher	117
	Associate	222
Programming languages	Associate	171
Statistics	Associate	117
Total	*Teacher*	*491*
	Associate	*2179*

Before any statistical tests, we decided to conduct Cronbach's alfa to check the consistency of the answers on the three questions regarding the assessment of teachers. It was of interest to see the Cronbach's alpha per subject per gender (Table 2).

Table 2: Cronbach's alfa per subject per gender of the assessed teacher

Subject	Gender of the teacher	N	Cronbach's alpha
Mathematics 2	F	50	0.612
	M	67	0.794
Discrete mathematical structures	F	65	0.723
	M	96	0.469
Introduction to information systems	F	25	0.160
	M	188	0.419

Taking the cutoff of 0.7 (Santos, 1990), we can see that only the scores of male teachers of Mathematics 2 and female teachers of Discrete mathematical structures have reliable scales. This means that the assessments are not consistent and that the SET should be inspected more closely.

To observe whether there is statistically significant difference between the scores male and female teachers and associates received we used Mann-Whitney test. Table 3 gives the results when it comes to the assessment of teachers based on the question "Compliance of the lecture and the scope of the subject". Firstly, it is important to observe the number of assessment per gender per subject. Accordingly, the first two analysed subjects (Mathematics 2 and Discrete mathematical structures) have an almost proportional number of evaluations per gender. On the other hand, the differences in the sample size per gender for the subject Introduction to information systems is high, so its results should be taken with caution. The test showed that there is statistically significant difference between male and female teachers, whereas female teachers received higher scores.

Table 3: Results of the Mann-Whitney test for the question "Compliance of the lecture and the scope of the subject"

Subject	Gender of the teacher	N	Mean	Z
Mathematics 2	F	50	4.58	-2.298*
	M	67	4.07	
Discrete mathematical structures	F	65	4.38	-2.144*
	M	96	4.28	
Introduction to information systems	F	25	5.00	-2.780**
	M	188	4.70	

Note: * $p<0.05$, ** $p<0.01$

The same test was run to explore whether there is difference when it comes to the suggested literature. Again, there are differences in all three subjects (Table 4). However, this time for subject Discrete mathematical structures the male teachers received higher grades; their mean grade was 4.77 compared to 4.32. In the other two subjects, female teachers are believed to have suggested higher quality literature.

Table 4: Results of the Mann-Whitney test for the question "Scope and quality of the suggested literature"

Subject	Gender of the teacher	N	Mean	Z
Mathematics 2	F	50	3.92	-3.092**
	M	67	3.30	
Discrete mathematical structures	F	65	4.32	-3.319**
	M	96	4.77	
Introduction to information systems	F	25	4.96	-5.244**
	M	188	3.39	

Note: ** $p<0.01$

Additionally, we wanted to inspect the overall impression the teachers left on the students at the end of semester. When it comes to mathematics related subjects, female teachers received better scores. The results for the subject Introduction to information systems showed that male and female teachers left the same impression on the students (Table 5).

Table 5: Results of the Mann-Whitney test for the question "Overall impression" of teachers

Subject	Gender of the teacher	N	Mean	Z
Mathematics 2	F	50	4.26	-3.999**
	M	67	3.60	
Discrete mathematical structures	F	65	4.63	-3.204**
	M	96	2.97	
Introduction to information systems	F	25	4.36	-0.444
	M	188	4.71	

Note: ** $p<0.01$

Finally, we observed the impression male and female associates lecturing eight different subjects left on students. The results were inconclusive (Table 5). In half of subjects, statistically significant difference was observed. In those four cases, twice male associates received higher grades (Statistics and Mathematics 2) and twice female associates (Decision theory and Business processes modelling).

Table 5: Results of the Mann-Whitney test for the question "Overall impression" of associates

Subject	Gender of the associate	N	Mean	Z
Statistics	F	65	4.42	-1.968*
	M	52	4.71	
Mathematics 2	F	34	4.59	-2.024*
	M	188	4.79	
Decision theory	F	127	4.87	-6.486**
	M	231	4.13	
Programming languages	F	36	4.78	-0.964
	M	135	4.60	
Data base	F	211	4.72	-0.828
	M	186	4.80	
Introduction to information systems	F	235	4.76	-1.191
	M	179	4.68	
Financial management and accounting	F	75	4.60	-0.755
	M	114	4.39	
Business processes modelling	F	56	4.66	-2.334*
	M	255	4.37	

Note: * $p<0.05$, ** $p<0.01$

4 Future directions of the study

Future directions of the study can be threefold: improvement of the current SET, analysing the lecturer's background and teaching methods, and applying multivariate methods on the SET results. Firstly, the current SET conducted at the University of Belgrade could be altered as it does not have a question regarding the gender of the surveyed student and mostly has subjective questions regarding the impression the lecturer has left. Introducing new questions might provide additional information upon which in-depth analysis could be performed. For example, knowing the gender of the student it could be explored whether the students' gender effects the scores given to male and female lecturers. Also, more questions which are not related to teaching could be added to better observe the presence of gender bias towards the lecturers.

Second potential future direction of the study is towards including more information on the teachers' background and their teaching methods. Teachers' and associates' years of experience in teaching the particular subject could provide additional insight on the potential presence of gender bias. Namely, in their recent research, Berbegal-Mirabent, Mas-Machuca, and Marimon (2016) examined how does the teaching experience affect student satisfaction. Using structural equation modelling (SEM) they proved that previous teaching experience positively influences student

satisfaction. It would be of interest to see whether this accounts for male and female lecturers with the same teaching experience. On the other hand, a lecturer should be free to engage in teaching methods he/she feels most appropriate to their subject (Surgenor, 2013). However, if both male and female lecturers use the same methods the difference in their SET scores could be a result of gender bias. These two aspects could be more closely examined in the context to gender bias.

Finally, another direction of the study could be towards applying confirmatory factor analysis via structural equation modelling (SEM) (Zhao & Gallant, 2012). For example, two SEM models could be compared – one, created based on the results of female students, and the other, created based on the results of male students. Also, the same could be done for male and female lecturers. Such approaches could show which variables have different factor loadings depending on the gender of the respondent and the gender of the lecturer.

5 Conclusion

The level of attention the academic community has devoted to the research on SET is staggering (Valsan & Sproule, 2008). Academics have been mesmerised with the opportunity to analyse the scores they have been given by their students. So far there are several directions in the research on SET: on its validity (eg. Spooren, Brockx, & Mortelmans, 2013), on its use by the faculty administration (eg. Valsan & Sproule, 2008), and on the effects which can influence the final results (eg. Bachen, McLoughlin, & Garcia, 1999).

Although SETs have their disadvantages their measures of class effectiveness and teaching performance are essential (Pounder, 2007). They provide both teachers and the administration with valuable data on how the teaching process has unfolded during a semester. In this paper, we have attempted to analyse in-depth one of SET's drawbacks: the impact of the teacher's gender on SET results. Gender bias is an important deficiency of student evaluations tests (MacNell et al., 2015) which should be tackled. Herein we explored the results of the mandatory summer SET at the Faculty of Organizational Sciences, University of Belgrade.

Among the 66 subjects whose lecturers were assessed we chose nine subjects and analysed 491 teachers' and 2179 associates' assessments. The SET whose results we analysed is specific as it is more related to the teaching assessment, which is highly subjective. Therefore, we based our research on the least biased questions where we could identify gender effect on the SET scores. The results of the teachers' SETs showed that there is gender bias as there were statistically significant differences in three subjects. In most cases, the female teachers received higher grades. To widen our research, we also observed the overall

impression lecturers left on students. The results were inconclusive: on some subjects female lecturers received higher scores, on some male lecturers, and on some there was no difference. However, the main two questions related to the scope of the lecture and the suggested literature, which are the same as all lectures on the same subject use the same materials, showed that defences exist. Some of our recommendations to reduce the observed gender bias is to raise the gender awareness both in the classroom, among the teaching staff, and among the administration.

This study has several limitations that should be pointed out. We cannot guarantee that these patterns would occur at another Faculty or University or country where student's expectations and attitudes might be quite different. Larger, multi-institutional samples are needed to better understand the role of gender in SET scores. Also, it would be interesting to see whether there are differences in the results between the Faculties within the same University. As all Faculties at the University of Belgrade conduct the same biannual SET, such research could be possible in the near future.

The results of this study have provided us with the basis for future exploration of the effects of gender on the SET scores.

References

Aleamoni, L. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, *13*(2), 153–166. doi: 10.1023/A:1008168421283

Aleamoni, L.M. (1987). Typical faculty concerns about student evaluation of teaching. *New Directions for Teaching and Learning*, *1987*(31), 25-31. doi: 10.1002/tl.37219873105

Anderson, K. J., & Smith, G. (2005). Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Sciences*, *27*(2), 184-201. doi: 10.1177/0739986304273707

Bachen, C.M., McLoughlin, M.M. & Garcia, S.S. (1999). Assessing the role of gender in college students' evaluations of faculty. *Communication Education, 48*(3), 193-210.   doi: 10.1080/03634529909379169

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., ... & Shepard, L. A. (2010). Problems with the Use of Student Test Scores to Evaluate Teachers. EPI Briefing Paper# 278. *Economic Policy Institute*.

Basow, S. A. (2000a). Best and worst professors: Gender patterns in students' choices. *Sex roles*, *43*(5-6), 407-417. doi: 10.1023/A:1026655528055

Basow, S. A. (2000b). Gender dynamics in the classroom. In J. C. Chrisler, C. Golden, & P. D. Rozee, *Lectures on the psychology of women* (2nd ed.). New York: McGraw-Hill

Berbegal-Mirabent, J., Mas-Machuca, M., & Marimon, F. (2016). Is research mediating the relationship between teaching experience and student satisfaction?. *Studies in Higher Education*, Online first, 1-20. doi:10.1080/03075079.2016.1201808

Bingimlas, K. A. (2009). Barriers to the successful integration of ICT in teaching and learning environments: A review of the literature. *Eurasia Journal of Mathematics, Science & Technology Education, 5*(3), 235-245. ISSN: 1305-8223

Feldman, K. (1993). College students' views of male and female college teachers: Part II—Evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, *34*(2), 151–211. doi: 10.1007/BF00992161

Galbraith, C., Merrill, G., & Kline, D. (2012). Are student evaluations of teaching effectiveness valid for measuring student outcomes in business related classes? A neural network and Bayesian analyses. *Research in Higher Education, 53*(3), 353–374. doi:10.1007/s11162-011-9229-0.

Hampton, S. E., & Reiser, R. A. (2004). Effects of a theory-based feedback and consultation process on instruction and learning in college classrooms. *Research in Higher Education, 45*(5), 497–527. doi: 10.1023/B:RIHE.0000032326.00426.d5

Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education, 5*(4), 419–434. doi: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. doi: 10.1080/13562517.2010.491911

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

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. doi: 10.1007/s10755-014-9313-4

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In *The scholarship of teaching and learning in higher education: An evidence-based perspective* (319-383). Springer Netherlands.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: a

multidimensional perspective. In: J. C. Smart (Ed.) *Higher Education: Handbook of Theory and Research*, Vol. 8, pp. 143-234.

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

Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education*, *27*(2), 187-198. doi: 10.1080/02602930220128751

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

Santos, J. R. A. (1999). Cronbach's alpha: A tool for assessing the reliability of scales. *Journal of extension*, *37*(2), 1-5.

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. doi: 10.1080/713611436

Spiller, D., & Ferguson, P. B. (2011). Student evaluations: do lecturers value them and use them to engage with student learning needs?.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching the state of the art. *Review of Educational Research, 83*(4), 598-642. doi: 10.3102/0034654313496870

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. doi: 10.1007/s11199-005-8292-4

Subramanya, S. R. (2014). Toward a more effective and useful end-of-course evaluation scheme. *Journal of Research in Innovative Teaching, 7*(1), 143-157.

Surgenor, P. W. G. (2013). Obstacles and opportunities: addressing the growing pains of summative student evaluation of teaching. *Assessment & Evaluation in Higher Education*, *38*(3), 363-376. doi:10.1080/02602938.2011.635247

Swartz, C. W., White, K. P., Stuck, G. B., & Patterson, T. (1990). The factorial structure of the North Carolina teaching performance appraisal instrument. *Educational and Psychological Measurement*, *50*(1), 175-182. doi: 10.1177/0013164490501021

Valian, V. (1998). Why so slow? The advancement of women. Cambridge, MA: MIT Press.

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. doi: 10.1080/00213624.2008.11507197

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. doi: 10.1080/02602938.2010.523819