DIFFERENCES IN PERCEPTIONS OF EFFECTIVE TEACHING BY FACULTY MEMBER'S GENDER AND STUDENT CHARACTERISTICS IN A TURKISH PUBLIC UNIVERSITY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF SOCIAL SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ECE YILMAZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE OF EDUCATIONAL ADMINISTRATION AND PLANNING
IN
THE DEPARTMENT OF EDUCATIONAL SCIENCES

SEPTEMBER 2018

Approval of the Graduate School of Social Sciences

Prof. Dr. Tülin Gençöz
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Prof. Dr. Cennet Engin Demir
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Serap Emil
Supervisor

Examining Committee Members

Assoc. Prof. Dr. Yaşar Kondakçı	(METU, EDS)	
Assist. Prof. Dr. Serap Emil	(METU, EDS)	
Assist. Prof. Dr. Burcu Tibet	(Ufuk Uni., EBB)	

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Ece Yılmaz

Signature:

ABSTRACT

DIFFERENCES IN PERCEPTIONS OF EFFECTIVE TEACHING BY FACULTY MEMBER'S GENDER AND STUDENT CHARACTERISTICS IN A TURKISH PUBLIC UNIVERSITY

Yılmaz, Ece
M.S., Department of Educational Sciences
Supervisor: Assist. Prof. Dr. Serap Emil

September 2018, 147 pages

The study aims to examine if there are statistically significant differences in student evaluations of teaching effectiveness by faculty members' gender and student characteristics including students' gender, grade and discipline. Data were gathered from 667 students studying in Middle East Technical University, Ankara. Data collection instrument had two parts: Demographic Information and Instructor Rating Questionnaire. Students were expected to select one of the faculty members whose course they had completed as the most or least effective, and rate how effective their teaching was. The first research question investigated if student evaluations of teaching effectiveness scores statistically differ by faculty member's gender and student gender. The results did not indicate a statistically significant interaction between faculty members' gender and student gender. Therefore, the main effects of faculty members' gender and student gender were investigated; however, the results indicated no significant main effect for faculty members' gender and student gender. The second research question investigated if student evaluations of teaching effectiveness scores statistically differ by the students' grade. The results indicated

that student evaluation scores significantly differ by students' grade. Lastly, the third research question investigated if student evaluations of teaching effectiveness scores statistically differ by students' discipline. The results displayed that student evaluation scores significantly differ by students' discipline. All in all, the results showed that there are not statistically significant differences in student evaluations of teaching effectiveness scores by faculty members' and students' gender; but there are statistically significant differences in them by student characteristics which are students' grade and discipline.

Keywords: student evaluations of teaching effectiveness, gender, grade, discipline, teaching quality in higher education

ÖZ

TÜRKİYE'DE BİR DEVLET ÜNİVERSİTESİNDE ÖĞRETİM ELEMANININ CİNSİYETİ VE ÖĞRENCİ KARAKTERİSTİKLERİNE GÖRE ÖĞRETİM ETKİLİLİĞİNİN ALGILANMASINDAKİ FARKLILIKLAR

Yılmaz, Ece
Yüksek Lisans, Eğitim Bilimleri Bölümü
Tez Yöneticisi: Dr. Öğr. Üyesi Serap Emil

Eylül 2018, 147 sayfa

Bu çalışmanın amacı öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinde öğretim elemanının cinsiyeti ve öğrencinin cinsiyeti, öğretim elemanından aldığı notu ve çalışma alanını içeren öğrenci karakteristiklerine göre anlamlı farklılıklar olup olmadığının incelenmesidir. Çalışmadaki veriler Ankara'da Orta Doğu Teknik Üniversitesi'nde öğrenim görmekte olan 667 öğrenciden toplanmıştır. Veri toplama aracı ise iki kısımdan oluşmaktadır: Demografik Bilgiler ve Öğretim Elemanı Değerlendirme Anketi. Öğretim Elemanı Değerlendirme Anketini doldurmadan önce öğrencilerden o zamana kadar ders almış oldukları öğretim elemanlarından en etkili ya da etkisiz olduğunu düşündükleri bir öğretim elemanını seçmeleri ve öğretim etkililiğini değerlendirmeleri istenmiştir. Çalışmanın birinci araştırma sorusu öğrencilerin öğretim etkililiği değerlendirmelerinin öğretim elemanının cinsiyetine ve öğrencinin cinsiyetine göre farklılaşıp farklılaşmadığını incelemiştir. Sonuçlar, öğretim elemanlarının öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinde öğretim elemanının cinsiyeti ve öğrencinin cinsiyeti arasında bir etkileşim ortaya koymamıştır. Etkileşim olmadığı için ana etkilere bakılmış,

ancak sonuçlar ne öğretim elemanının cinsiyeti ne de öğrencinin cinsiyeti için bir ana etkinin varlığını da ortaya çıkarmamıştır. Araştırmanın ikinci sorusu öğretim etkililiği değerlendirilen öğretim elemanının dersinden alınan notuna göre öğrenci değerlendirmelerinde sonuçlarında bir farklılık olup olmadığını incelemiştir. Buna göre öğrenciler tarafından öğretim etkililiğinin değerlendirmesinde öğretim elemanından alınan nota göre anlamlı farklar bulunmuştur. Son olarak araştırma, öğrencinin çalışma alanına göre öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinde farklılıklar olup olmadığını incelemiş ve öğrencinin çalışma alanına göre öğretim etkililiğinin değerlendirilmesinde anlamlı farklılıklar bulunmuştur. Özet olarak, çalışmanın sonuçları öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinde öğretim elemanın cinsiyeti ve öğrencinin cinsiyetine göre anlamlı farklılıklar olmadığını, öğretim elemanının dersinden alınan nota ve öğrencinin çalışma alanına göre ise anlamlı farklılıklar olduğunu ortaya çıkarmıştır.

Anahtar Kelimeler: öğrenci değerlendirmeleri, cinsiyet, not, çalışma alanı, yükseköğretimde öğretim kalitesi

To Aysel Yılmaz, the greatest mother a person can have

ACKNOWLEDGEMENTS

Firstly, I would like to express my gratitude to my advisor Assist. Prof. Dr. Serap Emil, who was always available to help me during my study. Her encouragement and guidance always motivated me while writing this thesis, and I am grateful to her.

Secondly, I would like to thank Assoc. Prof. Dr. Yaşar Kondakçı for his help and guidance throughout my Master's study. His valuable comments in my thesis defence also helped me enhance my thesis and research. I would also like to thank Assist. Prof. Dr. Burcu Tibet Arığ for her valuable comments and contribution to my study during my thesis defence.

I would also like to thank Sevgi Kaya-Kaşıkçı for her patience and time she gave me during the data analysis of this study. I learnt a lot from her, therefore, I am thankful to her. I am also grateful to Yusuf İkbal Oldaç, who was always ready to help me during my study.

I would like to thank my mother Aysel Yılmaz who has always supported and encouraged me to do my best. I am thankful to her for always believing in me.

I also like to thank my dear friend Peyman Yüksel who was always there for me during my study. I am grateful to her for her friendship, sharing her experiences with me and for providing me with a good environment to study. I am also thankful to my friend Özkan Özkara for his help during the writing process.

Lastly, I would like to thank to the Middle East Technical University academic personnel who allowed me in their classrooms to collect data and to all participants of the study who voluntarily helped me in the data collection process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This chapter of the study aims to introduce the main questions and discussions of the study and consists of four parts: The first part presents background to the study, and the theoretical frame for the study is discussed in this part. In the second part, the purpose of the study is given to offer the audience a better understanding of why this research is conducted. The third part of this section presents research questions, and related hypotheses are delivered in part four. With the aim of providing a rationale to the conduct of the present study, the significance of the study is discussed in part five. Lastly, in part six, operational definitions of the variables that shape the study are given.

1.1 Background to the Study

Because student evaluations of teaching effectiveness are used as a technique for faculty performance evaluation with both formative and summative purposes in some of the higher education institutions in Turkey, an investigation of various possible bias intertwined in their implementation, results, and interpretation becomes essential. To this end, this study aims to reveal if student evaluations of teaching effectiveness scores of faculty members differ by faculty member's gender and student characteristics.

Before dealing with the possible bias intertwined in student evaluations of teaching effectiveness, it is important to understand the context that has led to a more common use of student evaluations of teaching effectiveness as a faculty performance evaluation. Firstly, according to Harvey (2011), today with the increasing number of universities in the countries and the pressure to ensure quality and attract more students, a new consumerism, higher education consumerism, is in our lives. Students

are assigned the role of key stakeholders in this new trend. Although concepts like student course evaluations have long prevailed in literature, recently gathering feedback from students has been more significant in a world with enlargement of the higher education sector (Harvey, 2011). Secondly, Seldin, Miller and Seldin (2010) argue that increasing amounts of tuition fees students and their parents face have raised questions related to the quality of teaching in higher education institutions. In addition, Seldin, Miller and Seldin also argue that limitations on budget have forced universities to deeply examine each faculty member's teaching effectiveness. Thirdly, Lin et al. (2017) states that universities aim to conduct total quality management processes and improve the quality of education to attract more students. They also found that total quality management processes and teaching quality have critical and positive impact on student learning and loyalty; and that teaching quality intermediates the relationship between student learning satisfaction and student loyalty. Fourthly, Elez (2017) portrays universities as companies whose ultimate goal is to have satisfied customers, namely students. When students are satisfied, they stay in the institution, contribute to the positive representation of the university and attract new students. Marketing is necessary for the normal functioning of these institutions. This satisfaction depends on the quality of teaching, learning and working conditions, campus facilities and the environment in general (Elez, 2017). Lastly, Lavigne and Good (2014) argue that the mounting use of student evaluations as an indicator of teaching effectiveness has been due to an increasing interest in teacher influences on student outcomes, a larger prominence on teachers which holds them responsible for student achievement.

All these worldwide developments leading to a more common use of student evaluations of teaching effectiveness have also had some consequences for Turkish higher education and its universities. Balyer and Gündüz (2011) agree that together with the neoliberal policies, a lot of reforms and transformations which change and redefine the roles of universities have been taking place at universities in Turkey. Just like the worldwide trend, increase in the number of universities have created a competitive environment for Turkish higher education institutions. Alkoç (2017)

highlights the fact that the increase in the number of higher education institutions in Turkey has increased the number of alternatives for students to make a choice. In this competitive environment, universities pay attention to the issues like improving the quality of services and facilities offered to students to ensure student loyalty. In her study, Alkoç has found that student satisfaction, the image of the university and its curriculum are determiners of student loyalty. She suggests that university managements should place emphasis on to improve the quality of education by hiring faculty members of high quality, and increasing the quality of academic programs, products and services. Similarly, Baykal et al. (2002) also claim that universities make many decisions about curriculum, infrastructure and improving the quality of the instructors; and they position students as customers of these services while underlining the importance of determining student satisfaction and collecting data for quality improvement in higher education.

Middle East Technical University (METU) in Ankara is one of these universities which implement a student evaluation instrument at the end of each academic semester. Course Evaluation System at METU is an online evaluation of courses offered by faculty members of the university. The evaluation results are pursued by authorized personnel on the basis of university, faculty and department (METU Student Handbook, 2012). These evaluation results are used to award faculty members on the basis of their teaching by METU Mustafa Parlar Foundation. METU The Educator of the Year Award is given to those who have been found the most successful as a result of this Course Evaluation Questionnaire implemented by Rectorate. In the selection process, the means of both the last semester and cumulative scores of previous semesters are considered for the questions about the faculty member. Among the faculty members who satisfy the minimum criterion, the selection is done on the basis of other criteria which include being a full-time faculty member, the total number of courses offered during the academic year, the number of students and the number of evaluation forms filled in by students; and the selected faculty members are notified to the Administrative Board of Mustafa Parlar Foundation by METU Rectorate (METU Parlar Foundation, 2018). The results of the

Course Evaluation Questionnaire are also used as criteria for appointment and promotion for faculty members. In addition, METU Directorate of Personnel Department announces that faculty members are required to receive ratings higher than 4.00 for the last six semesters, or receive normalized ratings which are in the top 80 percentile to be appointed or promoted for the associate professor and professor positions in Natural Sciences, Engineering, Architecture, Social and Administrative Sciences. However, when these criteria are not satisfied, faculty members can use extra points from their publications (METU, 2018).

As can be seen student evaluations can be used as a way to measure faculty performance and reward faculty members at Middle East Technical University. However, the validity of these instruments is questioned in the literature (Spooren, Brockx, & Mortelman, 2013). For example, Onwuegbuzie, Daniel and Collins (2009) exploited a meta-validity model to run a meta-analysis investigation of student evaluations of teaching, and they concluded that there are not enough proof for content-related and construct related validity and according to them, this threatens the score-validity and usefulness of student evaluations. Similarly, Onwuegbuzie et al. (2007) analysed how students perceive the characteristics of effectiveness to investigate the validity of student evaluations of teaching with the participation of 912 students from different academic disciplines at a public university. The results of their analysis revealed that two of the canonical correlations showed a relationship between student gender among other student attributes and effectiveness themes, which meant a threat to content and construct-related validity of student evaluations. This study mainly focuses on faculty member's gender, and student characteristics which include student's gender, course grade and discipline to reveal possible bias in these student evaluation processes. Therefore, the next parts concentrate on gender and student evaluations of teaching effectiveness, and student characteristics and student evaluations of teaching effectiveness separately.

1.1.1 Gender and student evaluations of teaching effectiveness

One of the purposes of this study is to investigate gender differences in student evaluations of teaching effectiveness scores. Das and Das (2001) studied gender in student evaluations and examined the relationship between the gender and the gender role of the students and their best college instructor. The results of indicated that male business students tended to select a male instructor as their best instructor when compared to female students who tended to select females as their best instructor. The study further demonstrated that male instructors with low femininity scores and female instructors with gender-neutral characteristics were favored by their students more (Das & Das, 2001). As can be seen from this study, student evaluations of teaching effectiveness can differ by student's gender and faculty member's gender at the same time. Therefore, although student's gender is handled under the student characteristics variable in the title, the present study will handle both faculty member's gender and student's gender under the same research question.

To have a better and deeper understanding of gender issues in student evaluations as a faculty performance measure, it is critical to understand the gendered nature of higher education in Turkey. Öztan and Doğan (2015) claim that although the number of academic women is high enough to be worthy of admiration even in Western countries, the functioning of universities and male-dominated organization of work life in Turkey require a closer examination of this numerical visibility. For example, when the World Bank data for the percentage of female faculty members in Turkey are analysed, it can be seen that it was 22,73% in 1971; 24,6% in 1981; 31,7% in 1991, and 42,78% in 2014. At this stage, when we compare these percentages to the 2014 women in science regional data worldwide provided by UNESCO (2017), it can clearly be seen that the percentage of female faculty members is quite high in Turkey compared to the regional and world averages.

However, focusing only on the percentage of female faculty members in Turkey can be misleading because a deeper analysis of the numbers demonstrates gendered

patterns (Table 1.2.). One explanation for this numerical presence comes from Akbulut (2011). Akbulut mentions that academic women in Turkey created a space for themselves at Turkish universities with their academic identities; however, another dimension of this reality is that men prefer to work in private sector with higher wages while women continue to stay in academia with lower wages when their education levels are considered.

Table 1.1.

The Regional Averages for the Share of Female Researchers in 2014

Region	Percentage
For World	28,8%
For Arab States	39,9%
For Central and Eastern Europe	39.6%
For Central Asia	47.2%
For East Asia and the Pacific	22.9%
For Latin America and the Caribbean	44.7%
For North America and Western Europe	32.2%
For South and West Asia	19.0%
For Sub-Saharan Africa	30.4%

Note: Based on available data only. UNESCO (2017).

Another aspect of deeper analysis originates from the examination of the places and positions where the female academic staff accumulates. Acar (1996) draws attention to the accumulation of women in the positions of instructor, language instructor and assistantship in her article in which she shares the statistics of the year 1989. She underlines the fact that these positions have a more radical status by their job definitions when compared to the other academic positions. She considers this accumulation as false representation/participation on the grounds that these positions are low-level and fixed positions; in other words, closed to career advancement. Although Acar (1996) draws these conclusions depending on the statistics of the year 1989, today's distribution of male and female faculty members according to their academic ranks does not display a pronounced discrepancy. As exhibited in Table

1.2., especially for the highest positions of academic ranks e.g. professor and associate professor, the number of male faculty members is twice as many as the number of the female ones. Likewise, the number of male assistant professors is dramatically more than the number of female assistant professors. Nonetheless, the numbers get close to each other starting with instructor and research assistant positions, and the number of female language instructors doubles the number of male language instructors. Taking these points into consideration, it is obvious that as one proceeds toward the higher positions of the academic rank, women become less visible.

Table 1.2.

The Distribution of Male and Female Faculty Members in Turkey

Academic Title	Gender		Total
	Male	Female	
Professor	16734	7544	24278
Associate Professor	8839	5538	14377
Assistant Professor	21264	15674	36938
Lecturer (Öğretim Görevlisi)	11534	9073	20607
Instructor (Okutman)	3690	6283	9973
Research Assistant	22624	22992	45616
Lecturer, PhD	349	345	694
Total	85034	67449	152483

Note: Summary Number of Instructors. Retrieved from https://istatistik.yok.gov.tr on March 7, 2018.

Furthermore, the reflections of the previous argument can also be observed in the distribution of educational management positions among male and female academic staff. Among the 111 public universities in Turkey, solely 3 universities have a female President. This means that only 2.7% of universities have a female President. Similarly, in 104 universities which publishes the names of Vice Presidents on their websites, it is mentioned that there are only 27 female Vice Presidents among 228, which equals to 9.28% of all Vice Presidents (Habertürk, 2018). In a metaphorical

analysis study which analyzed female faculty members about themselves, Başarır and Sarı (2015) found that regarding women's invisibility, the metaphor used by women was "display decoration" to signal not only their numerical existence as something that is wanted but also to signal their absence in decision making processes and ignored existence.

It can be concluded from the discussions above that Turkish universities are "gendered institutions" in Acker's (1992) terms. According to Acker, gendered institutions are the structures traditionally established by men, currently prevailed by men, portrayed from a male point of view in superior positions and defined by nonexistence of women both currently and historically. In addition, Acker (2009) considers inequality in organizations as systematized unfair differences among participants in terms of power and control over goals, resources and outcomes; decision-making processes and practices; chances and possibilities for promotion and interesting work; job security and benefits; pay and reward systems regarding pay; and in esteem and job satisfaction. Acker (2009) adds that body differences between embodied male and female worker give hints about the appropriate assumptions which are pursued by gender-appropriate acts; and women and men in the workplace regularly use these gender-based assumptions along with the assumptions related to class and race while interacting with one another, which also has consequences in job evaluation processes which are also gendered as mentioned by Acker (1990).

Consequently, seeing that universities are gendered institutions, it is possible to argue that all these unfair differences have their own reflections in Turkish higher education and its job performance evaluation processes. Therefore, it is also important to investigate possible gender bias in student evaluations of teaching effectiveness instruments, their implementation and interpretation of results. For example, in a study which employed data from a French university, Boring (2016) found that although students seem to learn from both male and female professors equally, both male and female students acknowledged their male professors as more skilful in classroom leadership and more knowledgeable when compared to how they perceive

female professors. As is clear from this study, gender stereotypes and the distinct aspects of teaching are not independent of each other. The rationale underlying this is that students as the main information providers of these instruments may carry gender stereotypes regarding how an effective instructor should be and they may carry these gender stereotypes to their answers in the instruments. When this happens, the practice of student evaluations of teaching effectiveness may create difference and inequality among faculty members, as in Acker's (2009) terms.

1.1.2 Student characteristics and student evaluations of teaching effectiveness

The focus of student characteristics in this part are student's course grade and discipline. Firstly, literature presents a discussion of grade bias in student evaluations of teaching effectiveness in several different ways. As to the grade knowledge bias, Lin (2015) found a statistically significant correlation between students' grade and student evaluations of teaching effectiveness (SET) scores when students had midterms and received grade feedback before the implementation of SET. However, when grade knowledge was not available to students before the implementation, Lin did not find an association between students' course grade and SET scores. Therefore, if SETs are to be used both for formative and summative purposes as in the case of Middle East Technical University, the faculty members and administrators of the instrument need to be aware of such possible bias like gender and grade knowledge and take necessary precautions; otherwise, the results may be misleading.

In addition, another grade-related bias is studied by Isely and Singh (2005) by investigating the relationship between students' expected grades and student evaluations of teaching scores. They found that when a faculty had a course where students expect higher grades, the average student evaluation score in this course was more favourable. However, the authors explain it with the gap between students' expected grade and their cumulative grade point average rather than only expected grade.

Another study by Tata (1999) examined the relationships between fairness of grade distributions, fairness of the whole grading procedures and student evaluations of teaching. Tata revealed that in the cases where the actual grades failed to meet student expectations of grade distributions, in other words, where they were perceived to be unfair, grading procedures had an impact on student evaluations. In addition, it was also found that when students received their expected grade they gave higher ratings in student evaluations of teaching compared to those who did not get their expected grade.

Matos-Díaz and Ragan Jr. (2010) analyzed the student evaluations questionnaires at the University of Puerto Rico at Bayamón to investigate the impacts of grade distribution on student evaluations of teaching. The results of the study found showed that there was a negative correlation between student evaluation scores and the variance of expected grades. According to Matos-Díaz and Ragan Jr., this implies that faculty members can increase their scores by keeping the expected grade variance narrow. The results also supported the hypothesis that weaker students place a greater importance on this narrow distribution.

Braga, Paccagnella and Pellizzari (2014) used administrative data from Bocconi University to compare measures of teaching effectiveness to student evaluations of teaching. The follow-up performance of students was the estimation for effectiveness. The results of the study displayed that the classes in which higher grades were distributed were evaluated positively by students while the classes which are related to higher grades in future courses were evaluated negatively by students.

Consequently, if higher scores, narrower grade variation, students' perception of fairness matter in student evaluations of teaching effectiveness, an investigation of grade bias in these evaluations is of the great importance. One reason for this is put forward by Holmes (1972). Holmes studied the influences of grades and disaffirmed grade expectations on student evaluations of teaching effectiveness. In this study, half of the students who deserved A and B grades were given their expected grade, while

half of them were given a one-level lower grade. The results of the study indicated that when students' grade expectations are disaffirmed by the actual grade, students have a tendency to give lower ratings to their instructors not only in grading dimension but also in other dimensions of teaching effectiveness. When all these are considered, it becomes necessary to reveal possible grade bias in student evaluations of teaching effectiveness.

Secondly, another subject of bias in student evaluations of teaching effectiveness is discipline. Presenting a review of features of characteristics of teaching scholarship, Healey (2000) discussed the reasons to have diverse scholarships of teaching across disciplines in higher education. It was concluded that good teaching depends on context, is laborious and consists of more than one dimension. It was also highlighted that discipline-based education networks have critical importance in easing the communication and in the development of teaching scholarship. Parpala, Lindblom-Ylänne and Rytkönen (2011) studied how students conceptualized teaching quality in three disciplines. Their sample consisted of 695 students from the Faculties of Behavioral Sciences, Law and Veterinary Medicine. 12 of 21 dimensions of teaching quality revealed by both qualitative and quantitative analyses differed among the disciplines. The dimensions "Interaction", "Matching with students' prior knowledge" and "Encouraging critical thinking" were reported by the students of behavioral sciences more often while "Group work" and "Good materials" were reported more often by students of law. The students of the veterinary medicine, on the other hand, reported combining theory and practice more than students of law and they wanted more teacher support and opportunity to ask questions compared to other two disciplines. Parpala, Lindblom-Ylänne and Rytkönen (2011) concluded that there are differences in how students of different disciplines conceptualize teaching quality and this should be taken into consideration in the implementation of student evaluations of teaching at universities.

The way teachers conceptualize teaching in different disciplines is also important. Lindblom-Ylänne, Trigwell, Nevgi and Ashwin (2006) explored the relationship

between academic discipline and approaches to teaching, and the impacts of teaching context on approaches to teaching. The findings of their study revealed systematic differences in approaches to teaching both across disciplines and teaching contexts. This meant that teachers who encounter different teaching environments can have different teaching approaches in these environments. As to approaches to teaching, teachers of "hard" disciplines reported a more teacher-focused approach compared to the teachers of "soft" disciplines who reported student-focused approaches more. Similarly, Norton et al. (2005) investigated teachers' intentions and belief related to teaching in four institutions in the United Kingdom. The data came from 638 sets of responses through a questionnaire which aimed to measure nine aspects of teaching. The way teachers conceptualized teaching was seemingly the reason for the differences in teachers' intentions across disciplines and for the differences between men and women. On the other hand, differences in teacher's intentions in different institutions and differences among teachers who had different levels of teaching experience were seemingly a consequence of differences in the teaching context. Therefore, it was concluded that the way teachers conceptualize teaching and the teaching context together shape teaching intentions.

Barnes and Barnes (1993) investigated student evaluations of teaching in terms of generalizability and decision making. Six dimensions of teaching which include "organization, breadth of coverage, group interaction, enthusiasm, grading and individual rapport" constituted the evaluation items for two analyses. In one of these analyses, the object measurement was the teacher and in the second, it was the course. Although the findings indicated reliable decisions about the teachers with six dimensions, this was not true for courses. The reliability for course decisions differed among evaluation dimensions, and a similar pattern was found across disciplines. Therefore, Barnes and Barnes conclude that only one evaluation instrument may not be appropriate to be used in all academic disciplines.

When all this research about teaching and learning in different disciplines and the METU context are considered together, discipline-related bias in student evaluations

of teaching effectiveness also requires a close examination. Middle East Technical University has 5 faculties which have 41 different programs for undergraduate students; 5 graduate schools which have 105 masters and 70 doctorate programs, and a School of Languages (METU, 2018). When this wide variety of programs that the university has is considered, as Barnes and Barnes (1993) concludes, any discipline-bias may signal a need for discipline specific questions in student evaluations of teaching effectiveness.

All in all, taking into consideration concerns mentioned in the existing literature about the implementation and interpretation of student evaluations of teaching effectiveness and the contexts leading to the widespread use of these instruments as a way to ensure teaching quality across universities in Turkey, it is quite vital to investigate possible gender, grade and discipline-related differences which may stem from bias interwoven in the design, conduct or interpretation of student evaluations in the Turkish context.

1.2 Purpose of the Study

The purpose of this study is to focus on one form of faculty performance evaluation which is the student evaluations of teaching effectiveness. As can be seen in the literature presented above, universities are gendered institutions and job performance evaluation concerning faculty members is not also free from gender bias. Considering these gender bias and other bias related to student evaluations of teaching effectiveness, it is significant to investigate differences in these evaluation scores by mentioned variables. As a consequence of this situation, the present study aims to understand if there are any statistically significant differences in student evaluations of faculty teaching effectiveness by faculty member's gender and by student characteristics which include student's gender, course grade, and discipline.

1.3 Research Questions

Are there any statistically significant differences in scores of student evaluations of teaching effectiveness by faculty member's gender and student characteristics?

Sub-questions:

1. Is there a statistically significant difference in student evaluations of teaching effectiveness scores by faculty member's gender and student's gender?
2. Is there a statistically significant difference in student evaluations of teaching effectiveness by student's grade?
3. Is there a statistically significant difference in student evaluations of teaching effectiveness by student's discipline?

1.4 Hypotheses

Main research question:

H_0: There are not any statistically significant differences in scores of student evaluations of teaching effectiveness by faculty member's gender and student characteristics.

H_1: There are statistically significant differences in scores of student evaluations of teaching effectiveness by faculty member's gender and student characteristics.

Sub-question 1:

H_0: There is not a statistically significant difference in student evaluations of teaching effectiveness scores by faculty member's gender and student's gender.

H_1: There is a statistically significant difference in student evaluations of teaching effectiveness scores by faculty member's gender and student's gender.

Sub-question 2:

H_0: There is not a statistically significant difference in student evaluations of teaching effectiveness by student's grade.

H_1: There is a statistically significant difference in student evaluations of teaching effectiveness by student's grade.

Sub-question 3:

H_0: There is not a statistically significant difference in student evaluations of teaching effectiveness by student's discipline.

H_1: There is a statistically significant difference in student evaluations of teaching effectiveness by student's discipline.

1.5 Significance of the Study

This study is significant because it fills a gap in the Turkish context by analysing differences in student evaluations of teaching effectiveness by faculty member's gender and student characteristics which are student's gender, course grade and discipline. Anderson and Miller (1997) claim that over-reliance on student evaluations of teaching can be problematic due to the reasons mentioned earlier. Phill and Rasmussen (2013) also concludes that there is a demand for course evaluation systems which permits administrative control and are sensitive about the issues of multiple objectives, learning, academic freedom, reliability and validity of student answers. In addition, Boring, Ottoboni and Stark (2016) argues that bias to student evaluations may also stem from class size, course format and difficulty, physical environment of the classroom among other variables. Therefore, it is clear from the previous research that there are many different sources of bias to student evaluations. The present study and similar studies serve the purpose of uncovering the invisible bias that reduce the validity of student evaluations of teaching effectiveness.

Secondly, the outcomes of this study will be beneficial for higher education administrators and policy-makers. As can be seen in the Middle East Technical University context, student evaluations of teaching effectiveness are used for summative purposes for award distribution and promotion decisions. However, as Hornstein (2017) points out, the use of student evaluations of teaching effectiveness for summative personnel decisions could be inappropriate and even illegal in some contexts due to the validity concerns. For example, Anderson and Miller (1997) claim that students may have different expectations from their male and female instructors owing to the stereotypical perspectives they carry to the class. If these stereotypical views are reflected in the results of the present study and similar studies, these evaluations can be a source of discrimination in the workplace. Therefore, the outcomes of this study will help to clarify some validity issues related to these evaluations and can help higher educator administrators and policy-makers make decisions to continue, stop or revise these processes by taking into consideration the bias involved.

In addition, faculty members and students will also benefit from the outcomes of this study. McKeachie (1996) states that more than 90 percent of faculty members at the University of Michigan receive excellent scores from their students. However, when these excellent-rated faculty members look at their results, half of them find that they are below the average, and this is disappointing and demotivating as to teaching for faculty members rather than being enthusiastic about it. Especially if there are bias involved in student evaluations of teaching effectiveness, the results will also be demotivating for most faculty members, and therefore the outcomes of this study are expected to highlight the fact that student evaluations of teaching effectiveness may be misleading. Moreover, the outcomes will also be beneficial for students and student learning. Love and Kotchen (2010) claim that putting more emphasis on student evaluations worsen the grade inflation and may even diminish faculty members' effort to increase teaching quality. By highlighting the bias involved in student evaluations of teaching effectiveness, this study also questions how much

these evaluations measure teaching effectiveness and what their role is in promoting teaching quality.

Lastly, this study will be beneficial to have a broader comprehension of teaching effectiveness. To exemplify, Neumann (2001) presents a review of literature on disciplinary differences in teaching and claims that still more explanations are necessary to have a deeper understanding of teaching across disciplines in higher education. According to Neumann, these explanations and research are necessary to revise and shape both national and institutional policies for more just and successful administration of higher education.

All in all, due to consumerism of higher education, universities are continuously trying to adjust dynamic environments and altering needs, and therefore evaluations continue to preserve their place at the heart of decision-making processes in such dynamic environments. The present study serves to uncover possible bias interwoven to these evaluations to increase the validity of results.

1.6 Operational Definitions

The operational definitions of variables mentioned throughout the study are provided below.

Faculty member's gender: This is an independent variable, it is categorical and nominal by nature and aims to determine if the faculty member who is evaluated by the student in this study is male or female. For the assessment, the participants are simply expected to state the gender of the faculty member they choose to evaluate.

Student's gender: This is also an independent variable, it is categorical and nominal by nature and aims to determine if the student who is evaluating a faculty member in this study is male or female. For the assessment, the participants are simply expected to state their own gender.

Grade: This is another independent variable which is treated as categorical and nominal in this study. It aims to determine the student's course grade which is received from the faculty member evaluated in the present study. For the assessment, the participants were simply asked to state the grade they got from the chosen faculty member's course. The grade variable has 12 levels in this study which are explained in detail below according to the information obtained from METU Directorate of Student Affairs (2018).

Table 1.3.

The Letter Grades, the Coefficients of the Letter Grades, the Score Intervals Used in the Calculation of Grade Point Average at METU

Letter Grade	Coefficient	Score intervals
AA	4,00	90-l00
BA	3,50	85-89
BB	3,00	80-84
CB	2,50	75-79
CC	2,00	70-74
DC	1,50	65-69
DD	1,00	60-64
FD	0,50	50-59
FF	0,00	0-49
NA	0,00	*

Grade NA is given to the students who did not fulfil the attendance requirements of a course, who did not fulfil course requirements to take the final examination, and who did not take midterm and final examinations. NA grade is treated as grade FF in mean grade calculations. Grade S (Successful) is given to the students who completed a non-credit course successfully, and it is not included in the calculation of Grade Point Average. Grade W (Withdraw) is given to the students who have withdrawn from a course.

Discipline: It is an independent variable which is categorical and nominal as to its characteristics. It aims to determine the discipline the students are studying in. For the assessment, the participants were simply asked to state the department they study, and departments were later grouped by the researcher depending on which discipline they belong. The discipline variable in this study has 6 levels and each includes the departments mentioned below.

1. Architecture: Architecture, City and Regional Planning
2. Humanities: Sociology, Psychology, History
3. Engineering: Environmental Engineering, Electrical-Electronics Engineering, Food Engineering, Industrial Engineering, Chemical Engineering, Geological Engineering, Civil Engineering
4. Natural Sciences: Physics, Mathematics, Molecular Biology and Genetics, Chemistry
5. Economics and Administrative Sciences: Business Administration, Political Science and Public Administration, Economics, International Relations
6. Educational Sciences: Physics Education, Elementary Science Education, Mathematics Education, Elementary Mathematics Education, Early Childhood Education, Foreign Language Education, Computer Education and Instructional Technology

Student evaluations of teaching effectiveness: This is the dependent variable of the study and assesses the teaching effectiveness of the faculty member who is chosen by the students through Instructor Rating Questionnaire. The dependent variable is continuous and interval as to its characteristics. The assessment is done on a 5-point rating scale (1=strongly disagree, 5=strongly agree, and 0= Not applicable). Some of the items of teaching effectiveness scale are: "The instructor communicated effectively", "The instructor had a good sense of humour" and "The instructor would motivate students to do their best".

CHAPTER 2

REVIEW OF THE LITERATURE

The literature review in this study consists of three main parts. The first part involves definitions and a discussion of faculty performance evaluation in higher education. In this part, different techniques of faculty performance evaluation are also introduced with a specific focus on student evaluations of teaching effectiveness. The second section focuses on the gendered nature of performance evaluation setting its connections to the gendered nature of faculty performance evaluation. The last part of this section provides some examples of previous studies in the literature about gender bias and student evaluations of teaching. Lastly, in the third part, a summary of discussions covered in this chapter is presented.

2.1 Faculty Performance Evaluation

As student evaluations of teaching effectiveness are used as a concrete form of faculty performance evaluation measures in practice in some higher education institutions, it is important to understand what faculty performance evaluation is, what purposes it serves, and its practical uses. Therefore, this part of the literature review aims to answer these questions. Arnautu and Panc (2015) defines performance evaluation as an organizational process that is used to assess staff in an ongoing and standardized way. The need for standardization and objectivity creates the need for explicit criteria in measurement. These criteria should be observable and standardized in an acceptable way, their definitions should be obvious (Arnăutu & Panc, 2015).

As is clear from the definition given above, performance evaluation requires a transparent measurement with observable and standardized criteria. As to the question of why this standardization has become important in higher education to

measure faculty performance, we can find some answers in Curry (2006). According to Curry (2006), such faculty performance reviews have gained importance recently for several reasons. Firstly, universities have started to use salary increases depending on the faculty performance, and this required a relatively concrete performance evaluation. Secondly, the alignment between the faculty and the institutional goals have gained significance in an environment which has limited resources. Thirdly, tenure and promotion decisions have also been under the impact of faculty performance evaluations. Due to the criticisms received from public, or as a result of institutional self-criticism, many universities have started to do post-tenure evaluations to make sure that tenured faculty members have acceptable performance levels (Curry, 2006).

As to the purposes that faculty performance evaluation measurement serves, it is important to understand Channing's 2016 research because it clearly shows that although it mainly serves two significant purposes as a form of summative and formative evaluation, these purposes may change depending on the whether one occupies a faculty member position or an administrative position. In this study which was conducted with the participation of 10 tenured faculty members and 12 administratives, Channing (2016) found that while the major reasons were stated as to foster development for educators (40%) and to assess overall performance (40%) as well as evaluating teaching performance (20%), the administrators stated that the main purpose was to evaluate teaching effectiveness (50%), to foster development for educators (42%) and to assess overall performance. This clearly shows that administrators link this information to different kinds of summative evaluation (Channing, 2016).

Kalaycı (2009) handles the issue from the perspective of Turkish Higher Education and claims that the quality assurance systems at universities aim to increase faculty research and teaching performance. If an institution wants to develop and reach its targets, it needs to evaluate staff performance. In many universities of the world, evaluations are conducted for summative and formative evaluation purposes.

However, there has not been an effective evaluation system which can include both purposes in a systematic way. Turkish system that uses these systems has also its problems. These evaluations lack meticulous work and mostly not used for its real purposes. In most universities, performance evaluation is limited to only student evaluations of teaching and they are regarded only as part of a routine work conducted due to quality assurance systems. If universities want to catch international standards, they need well-functioning performance evaluation systems for their faculty members (Kalaycı, 2009).

Consequently, as can be seen from the discussions presented above, faculty performance evaluation is an ongoing and standardized process, serves mainly two purposes as summative and formative evaluation. However, its practical uses are not free from several drawbacks such as lacking meticulous work and being used for more superficial purposes rather than its real purposes. Therefore, the search for better functioning performance evaluation systems should be what higher education institutions are focusing on. To be able to find out such a functioning system, it is important to have knowledge of different ways to evaluate faculty performance including peer reviews, teaching portfolios, and student evaluations of teaching.

2.1.1 Peer review of teaching

Peer review is one of the techniques used in faculty performance evaluation. Chism (2007) defines peer review of teaching as informed associate verdict about the quality of teaching either for professional development or staff decisions. Chism clarifies that informed verdict in this definition refers to a systematized work which is based on concrete and suitable proof and mental analysis. Similarly, an associate is a "knowledgeable judge". The definition of this informed associate, who is a peer by definition is further handled by Arreola (2007). Arreola mentions three different types of peers: an internal content peer, an external content peer, and an internal non-content peer. Accordingly, an internal content peer is a colleague in the same institution who has the same content knowledge as the person whose teaching is

assessed. In addition, an external content peer is a colleague who is not from the same institution, but still has the same content knowledge. An internal non-content peer is a colleague from the same institution; however, this peer does not have the same content knowledge as the person being evaluated.

As to the question of what purposes peer review of teaching serve, Keig and Waggoner (1995) mentions a difference between its summative and formative purposes. The mistaken assumption that was hold by faculty members that summative evaluation which is used to make decisions in reappointment, promotion and tenure processes could also be used to enhance instructional quality. However, this assumption was challenged and criticized and Keig and Waggoner (1995) highlights that summative evaluation hardly provides adequate information for the faculty members to help to improve teaching. What was suggested as an answer to this challenge was formative evaluation whose main purpose was to enhance instructional quality. Formative peer evaluation was the solution here, as a process in which faculty members work together to evaluate each other's teaching and help each other to improve teaching. It included "direct classroom observation, videotaping of classes, evaluation of course materials, an assessment of instructor evaluation of the academic work of students, and analysis of teaching portfolios" (Keig, & Waggoner, 1995).

Although its objectives are clear, there are some issues regarding how to use peer reviews of teaching. Since peer review of teaching aims at professional development, Yiend, Weller and Kinchin (2014) suggests that it should be "professionally guided" to be beneficial for the faculty. Their study is a single case study of an instructor who attended a single teaching observation cycle which was administered with an educational expert and peers from the same field of study who performed the roles of both an observer and observee. The study revealed that the contribution of peer review of teaching to faculty was evoking reflective approach to observation, and it was only possible when the faculty had enough comprehension of its objectives and administration. Therefore, a "professionally guided peer review" of teaching observation in which its objectives and administration were informed through formal

trainings or modelling of feedback and experience with specialists in education was suggested (Yiend, Weller, & Kinchin, 2014).

The arguments presented in Yiend, Weller and Kinchin's 2014 study were also supported by Barnard et al. (2015). They also maintained that it is not realistic to hold the expectation that everyone will accept and directly join the peer review process without a need for more help, guidance and leadership. Their suggestion for this drawback of peer review of teaching was distributive leadership which meant having empowered relations, a distinct vision and distributed accountability. They conclude that this distributive leadership is essential for the efficacy of peer-review process in the long term (Barnard et al., 2015).

As to the faculty perceptions of peer review of teaching and its benefits, a recent study by Lowder et al. (2017) revealed positive faculty attitudes. The study was conducted at Kennesaw State University, which offers "Teaching Partners Program" open to all faculty staff who teach at least one course during one semester. In the program mentioned above, faculty members from different departments are put in pairs, and they observe their teaching and deliver feedback. The survey comments of faculty members who participated in the program revealed some benefits which included discovering new pedagogic techniques, new partnership possibilities, beneficial proof for promotion and tenure and broadening their perspectives (Lowder et al., 2017).

Despite the benefits mentioned above, peer review of teaching has also its drawbacks. Some of the drawbacks which were listed by Berk (2006) are as follows: Peer reviews are subjective judgments which may turn out to be unreliable sources of information. There are also problems of impracticality in that classroom observations take too much time, and observations of one or two classroom hours are not representative enough to have a valid idea of the instructor's teaching performance. Moreover, there are problems stated regarding the peer review scales and the way they measure

teaching effectiveness, which all make peer reviews of teaching inappropriate data for summative decisions made by administrators in higher education.

To summarize the discussion above briefly, while peer review of teaching is can be used for both and summative purposes in higher education institutions, it is important to keep in mind that summative use does not necessarily allow an increase in the teaching quality. What makes peer reviews beneficial for the faculty to improve their teaching is formative use in a professionally guided way. Although faculty report positive attitudes toward peer review of teaching by claiming that it allows professional development opportunities like discovery of new pedagogical techniques and new partnership opportunities, this faculty performance evaluation technique has also its limitations. These limitations include but not limited to being subjective, impractical and inappropriate for summative decisions. As a consequence, higher education administrators should not ignore these limitations in its implementation.

2.1.2 Teaching portfolios

Teaching portfolios are another method used in the evaluations of faculty teaching effectiveness. Devanas (2006) defines teaching portfolios as a record of materials collected to portray a lecturer's teaching effectiveness. It is supposed to be around 8-10 pages and presented with sample course materials, student work, and evaluations of teaching effectiveness including but not limited to student evaluations of teaching or peer reviews of teaching. Devanas adds that the most important components of teaching portfolios are statements of teaching responsibilities, achievements, and reflections on the faculty member's own teaching philosophy. Chism (2007) states that teaching portfolios are part of summative and formative evaluations of teaching. The most popular summative type is the "promotion and tenure dossier", and in the formative type it serves to uncover the faculty member's philosophy and influence on teaching.

As can be seen in the previous paragraph, teaching portfolios are more comprehensive when compared to student evaluations and peer reviews only. Seldin, Miller and Seldin (2010) suggest that student evaluations and peer reviews resemble a torch in that they enlighten only a minor part of a whole teaching performance. However, in their opinion, teaching portfolios are like searchlights, and similar to these very bright lights, teaching portfolios uncover a wide scope for various teaching skills, capabilities, approaches, philosophies and methods.

As to the question of what purposes a teaching portfolio serves, although Devanas (2006) lists them as personal development, personnel decisions including hiring, award distribution and documentation of prolonged careers, Erikson, Erlandson and Erikson (2015) point out the limitations of using teaching portfolios for professional purposes. Their argument is that teaching portfolios may end in academic dishonesty which may include presenting false information regarding achievements; and these are difficult to detect due to vague and perplexing criteria teaching portfolios employ. The solutions they offer include redefining academic quality and agreeable academic conduct, redefining the ways teaching skills are characterized and evaluated, and lastly using teaching portfolios only for personal development purposes (Erikson, Erlandson, & Erikson, 2015).

In addition, some instructors mention some other drawbacks of writing teaching portfolios in the studies in teaching portfolio literature. One of these studies is done by Donaghue and Dolci (2013). Apart from a discussion of teaching portfolios as a means of professional development and making recommendations about the components of a professional teaching portfolio, their study also presents information about the use of teaching portfolios in teacher evaluation at a federal university in the United Arab Emirates. Although some teachers stated the benefits of teaching portfolios as collaboration with colleagues which leads to opportunities in learning from each other, and continuous reflection on the professional acts, some other faculty members regarded teaching portfolios as an additional workload. Some requirements of the process like updating the online database was found time-

consuming, confusing and demanding. Faculty members also find them worthless when they cannot do them immediately after they attend a professional development session because later it becomes more difficult to remember details. As a result of these drawbacks in the implementation of teaching portfolios, some faculty members believe that they write these portfolios just to make the management pleased (Donaghue & Dolci, 2013).

Despite the drawbacks mentioned in the previous paragraph, the success of teaching portfolios in presenting holistic information through multiple sources of evidence (Chism, 2007) and in presenting the philosophy behind teaching practice (Hamilton, 2017) is undeniable. Hamilton (2017) shares her own experience which is transition from being a teacher to a teacher educator, and she maintains that teaching portfolio assisted her in creating a teacher educator identity and gave her a philosophical awareness of teaching practice (Hamilton, 2017).

All in all, just like peer reviews of teaching, teaching portfolios are used for both summative and formative purposes. When compared to peer reviews and student evaluations of teaching; however, they are more comprehensive in that they give a philosophical awareness about one's teaching practice and help to create a teacher identity. Despite these positive contributions to faculty, teaching portfolios also come with their limitation in practice. Therefore, while using teaching portfolios, higher education administrators need to pay attention to these limitations.

2.1.3 Student evaluations of teaching effectiveness

When the definition of evaluation in an educational setting is taken into consideration, Kember and Ginns (2012) states that the most prevalent evaluation model used to measure faculty performance consists of teaching and course evaluation surveys. According to them, evaluation is defined as obtaining, examining and interpreting data in order to affirm decision-making. The uses of evaluation in higher education involves decision making processes regarding design, development,

revisions and actions which are teaching and learning related. The major purposes of evaluation in a teaching and learning context should be raising the teaching quality, ensuring that these actions occur under improved conditions, and creating the opportunity for learning outcomes. For these purposes, each term universities require students to fill in numerous surveys. Due to the widespread implementation, these surveys have started to be used as a synonym for "student evaluation" (Kember and Ginns, 2012).

Like other types of faculty performance evaluation, student evaluations of teaching are also used for both summative and formative purposes. An example for summative purposes comes from Universiti Teknologi Malaysia. Puteh and Habil (2011) states that Universiti Teknologi Malaysia gives annual Teaching Awards to two academics who performed successfully in both teaching and learning. The candidates are expected to receive scores higher than the minimum required score of the university for three sequential years. An example of formative use comes from Australian universities. Chalmers (2011) asserts that all Australian universities also have a student feedback system which is addressed primarily at the instructor to promote instructors' individual improvement. In addition to the individual-level use of results, institutions have started to use student feedback to inform performance and promotion decisions although student feedback was restricted to the individual use of the instructor in the past. Similarly, in Malaysia for instance, there are some universities which utilise student evaluations of faculty for formative purposes. Puteh and Habil (2011) addresses Universiti Putra Malaysia (UPM) and University Sains Malaysia (USM) as universities where student course evaluations are used to enhance the quality of instruction. At Universiti Putra Malaysia, the instructors who got scores lower than the minimum score demanded by the university are expected to participate in teacher training courses and similar activities. At University Sains Malaysia, on the other hand, proactive measures are given importance. To exemplify, if an instructor gets a low score, he/she is invited by their supervisors to talk about the reasons for the poor score.

Therefore, student evaluations are valuable as a component in the evaluation processes of teaching and learning contexts. Pallet (2006) argues that two important reasons for this value firstly come from the fact that students have the best chances to monitor teaching, and secondly, they are the target audience of teaching act. Students are good at evaluating the professor-student relationships and the exactness of communication. In addition, their views on workload, difficulty and grading and self-assessment of their own learning are quite related to teaching effectiveness (Pallet, 2006).

Despite these advantages mentioned in the previous paragraph, the benefits of student course evaluations in enhancing the quality of teaching and learning are still questioned and whether these surveys really serve the purpose of enhancing teaching and learning quality has not been proved yet (Kember and Ginns, 2012). One reason for little evidence of such benefits may stem from the consumerist idea that the center of student feedback is usually about teaching and facilities rather than learning (Harvey, 2011). Therefore, keeping these limitations of student course evaluations in mind, it is remarkably critical to work on their purpose, uses and misuses.

2.1.3.1 Teaching effectiveness

Since the goal of teaching and course evaluations is to improve the quality of teaching and to improve student learning, defining and measuring teaching effectiveness have become a critical issue in higher education. Hénard and Roseveare (2012) mention several contextual shifts which have led to changes in the higher education institutions regarding teaching quality. These contextual shifts include but not limited to the internationalization of higher education, the expanding capacity of education and growing diversity of students, rapid changes in technology which can rapidly make curriculum and teaching practices old-fashioned, the increasing need for graduates who can take civic responsibilities, the need for skills which will help universities to compete in the global arena, financial efficiency, and the need to bring up a talented workforce who can respond to the needs of the 21st century.

While these contextual changes have been taking place, universities have wanted to attune themselves to the altering needs of their stakeholders. Nasser-Abu Alhija (2016) claims that higher education institutions want to progress their teaching practices for several reasons. Initially, as these institutions serve more than one stakeholder with diverse expectations, they want to ensure that they are trustworthy sources of good teaching. Secondly, since both employers and policy makers expect universities to prepare students for satisfying employment and professional development, universities are required to act in response to these expectations. Thirdly, research productivity and quality alone are not any more adequate to retain reputation and a good performance regarding teaching and learning are indispensable aspects to have a good reputation (Nasser-Abu Alhija, 2016).

There are unavoidable impacts of these contextual changes on the definition of teaching effectiveness. However, in order to understand what teaching effectiveness means, we should firstly engage in the definition of teaching itself. Arreola (2007) defines teaching in a broad sense as performing a set of actions that involve interactions with students in order to ease, support and ensure student learning. Related to the definitions of teaching, Kerchner and Mitchell (1986) underline the fact that teaching is not a fixed and predesigned set of routine acts, but it necessitates an excessive level of teacher adaptability to various requirements of the teaching environment, which allows teachers to pay attention to individual student reaction while teaching at the same time. They further argue that good lessons resemble beautiful conversations in that they are both transformed by the responses they arouse (Kerchner & Mitchell, 1986).

Based on the facts underlined by Kerchner and Mitchell (1986), we can understand that not all teaching is effective. However, it is not an easy task to give a universally accepted definition of effective teaching for different reasons (Ryans, 1949; Arreola, 2007). To illustrate our sightlessness in describing effectiveness in teaching, Ryans (1949) gives the example of a blind man whose opinion of an elephant changes relying on which part of the animal's body he touches. Ryans maintains that we are

identically sightless as to defining teaching effectiveness, and most of us have different ideas about what makes effective teaching (Ryans, 1949).

As mentioned earlier, because a universally accepted definition of teaching effectiveness has not been agreed upon, the concept has also taken different forms in different time periods as Borich (2007) mentions. Borich states that a revolution which has taken place in the definition of good teaching in the last thirty years. Using social ideals as a criterion started to be found unrealistic, and teachers' psychological characteristics were not found to be a strong indicator of good teaching. Since then, the concept good teaching has been replaced by effective teaching, and the studies related to effective teaching stopped to focus only on teacher characteristics, and they have started to include their impacts on learners. In this way, the learner and the teacher-learner relationship have become the center of current definitions of effective teaching (Borich, 2007).

The main discussions of 1980s which have led to the changes mentioned above can also be find in the arguments put forward by Kerchner and Mitchell (1986). They suggest that the relationship established between the teachers and learners drastically affect student learning. Accordingly, immensely committed teachers are more aware of their students and their concerns, and they can canalize their concentration better on their students. The interruptions to their concentration may have serious effects especially on learners who can easily lose their attention and get lost. As a result, being able to reach learners is not a single matter of covering the classroom material, or presenting a lesson plan, but it is a matter of concentration and commitment.

Despite these discussions going on in the teaching effectiveness literature, studies allowed some researchers to define what teaching effectiveness is. Nilson (2010) defines teaching effectiveness as the level of achievement in promoting student learning. According to this definition, when students learn more and are involved in the desired learning outcomes at cognitive, affective, ethical and psychomotor levels,

they are better at revealing these outcomes, which is an indicator of teaching effectiveness.

As to other definitions of teaching effectiveness in the literature, van der Lans, van de Grift and van Veen (2018) suggests that effective teaching is a developmental process in that these teaching behaviors develop in a stage-wise manner. The initial stage in this process is providing a safe learning environment which is followed by behaviors for efficient classroom management and instruction quality. When they gain adequate skills in these domains, they start developing skills regarding teaching methods, teaching learning strategies and individualizing the course content to satisfy individual students' needs (van der Lans, van de Grift, & van Veen, 2018).

In Bidabadi et al.'s (2016) study, the teacher participants of the research defined teaching and learning as "a shared process", which assigns duties not only to teachers but also to students. Through this shared process, universities were expected to make students questio their already-existing beliefs and mindsets about the running of the world in order to make them an elevated level of conception (Bidabadi, Isfahani, Rouhollahi, & Khalili, 2016). As a result, because teaching is not a one-way linear act from teacher to student, an investigation of students' views about teaching effectiveness becomes vital.

Alemu (2014) gathered data from 69 engineering students and 64 engineering instructors to investigate practices regarding teaching effectiveness and effective instructor's characteristics in engineering education in the distinguished universities of Ethiopia. The analyses did not reveal a significant difference between how students and instructors perceive the characteristics of effective and ineffective teaching. Both instructors and students described the effective instructor as the one who could show respect, make the lessons interesting, treat students equally in assessment, give importance to student achievement, love the subject taught, be friendly, stimulate and support an environment in which discussions and questions arise, be well-prepared and organized, make complicated subjects easy to handle for students. Although the

findings of the study suggest that the instructors' personality is the leading factor for teaching effectiveness, this list suggests that both personality traits and teaching skills are essential for teaching effectiveness (Alemu, 2014).

Allan, Clarke and Jopling (2009) conducted a study to examine students' understanding of effective teaching at a modern university in the United Kingdom. In the study, effective teaching is portrayed as a matter of four domains: "providing a supportive learning environment; having high expectations; scaffolding learning; and providing clear explanations". The data were collected through a scale from 80 students, and 65 of these students participated in focus group discussions. The researchers conclude that an encouraging learning environment in which instructors scaffold effective student learning and stimulate effective communication with their students were found to be more essential for effective teaching when compared to having high academic expectations (Allan, Clarke, & Jopling, 2009). These findings also support the findings of Bidabadi et al.'s (2016) study which regarded teaching as a shared process. The findings of this research also maintain that students want to see themselves as partners in learning not as passive recipients of the teaching process. Therefore, it is concluded students themselves have a very important role in making teaching effective.

Lastly, Brown and Atkins (1988) define teaching effectiveness in relation to one's goals. Therefore, what can be regarded as effective in one teaching context may not be effective in another context. Devlin and Samarawickrema (2010) also highlights the importance of teaching context in defining effective teaching. According to them, this context is prone to uninterrupted and numerous transformations due to both inside and outside forces regarding universities. Consequently, they claim that the definition of effective teaching should change and develop in order to respond to and reflect the requirements of the context where teaching and learning occur (Devlin & Samarawickrema, 2010).

As can be seen from the discussions above, teaching effectiveness does not have a single valid definition and teaching is a responsive act shaped by the teaching context and its participants. Despite this fact, Seldin (2006) underlines that research reveals consistent results as to the traits of effective teaching, and coming to the class prepared for the lesson, having extensive knowledge about the subject matter taught, increasing student motivation, treating students equally without favoritism or discrimination, being able to deal with the details of learning well, eagerness both in teaching and the subject matter are among those traits. As to the use of student evaluations of teaching instrument to evaluate teaching effectiveness, Seldin adds that as long as the questions asked in student evaluation surveys are asked appropriately and properly, students are able to grant trustworthy and accurate information about how effective the course and the act of teaching are.

However, according to Spooren, Brockx and Mortelmans (2013), that different SETs which are used to evaluate teaching effectiveness in fact do not reflect a single accepted definition of good teaching can lead to problems in their practical use. They argue that different contexts require retesting the usefulness of the instrument in the new context. The example comes from the shift from teacher-centred to student-centred teaching. Moreover, the recent studies also question whether personal traits lead to bias in SETs or cause halo effects and underlining that they may have effects on the interpretation of results. That the conclusive idea of how the characteristics of students, courses, and teachers affect the results have not been reached although the instruments continue to be used widely. Therefore, the authors question the usefulness of SETs as both formative and summative purposes and underlines that they are still a controversial topic in higher education as research has not yet been able to provide explicit answers regarding the validity of these instruments (Spooren, Brockx, & Mortelmans, 2013).

All in all, the definitions of teaching and teaching effectiveness have changed over time due to other contextual changes, and although some part of research considers student evaluations of teaching instruments as a way of gathering information about

teaching effectiveness, other part of research still questions their uses. The next part presents a discussion of these issues that are questioned about student evaluations of teaching effectiveness.

2.1.3.2 Issues regarding student evaluations of teaching

Although students are supposed to provide valuable feedback about teaching effectiveness, research also showed that overreliance on student evaluations while measuring teaching effectiveness does not always produce desired outcomes. Firstly, Pallet (2006) states that students do not have the qualifications to judge some components of effective teaching. These include judging the suitability of professor's objectives, knowledge in the field, the quality of materials, and suitability of assessment techniques used in the course.

Another warning similar to Pallet (2006) comes from Morley (2014). Morley examined 24,295 student evaluations of teaching in 1280 sections of courses during the spring term in 2010 at the University of Colorado, Colorado Springs. The data originated from 67 departments and fields of study. The instrument used in this study was a nine-item SET instrument which was developed by the Chancellor's Advisory Committee on student evaluations at the University of Colorado. The author warns against the heavy reliance on SETs as students were found to be better at listing instructors' strong and weak points when compared to their ability to make absolute comments about the quality of instruction (Morley, 2014).

Another concern about overreliance on student evaluations originates from the response rates. Pallet (2006) highlights that when the percentage of students who answer the questions in the survey is lower than 65, the degree of representativeness the data carry can be questioned. In addition to representativeness, the reliability and validity of these instruments are also questioned in the literature. As to reliability issues, consistency, stability and generalizability of results are the main principles (Nilson, 2010). As to validity, it is about the question of whether course/instructor

evaluations can really measure what they are supposed to measure (Field, 2018). According to Nilson (2010), there are two questions related to validity of student evaluations. One is whether they can reflect student learning adequately and second is whether there are any biases that can affect the results. As student evaluations and student learning get close to each other, the biases involved get fewer and feebler. The following are some studies that question validity and reliability of student evaluations of teaching effectiveness instruments.

Padró (2011) mentions Seashoal Lows University, which is a teaching focused university in the United States with its 8000 undergraduate and graduate students, as a case study and criticizes uses of student evaluations of faculty as a key component of promotion and tenure decisions. The concerns mentioned here are similar to those mentioned above which are related to validity and reliability of the instruments. Faculty members think that the instrument is not directly related to the organizational criteria of effective teaching, students are not knowledgeable enough to evaluate content, which can turn the evaluation into a "popularity contest". In addition to these, the lack of instructional support for faculty to help to improve their teaching is another issue for critics.

In the interpretation of results of such surveys, Rotenberg (2005) claims that there is a possibility that students can make connections between the conditions that the instructor is accountable for and the conditions that she/he is not accountable for. To exemplify, a student who is going through hard times in his/her personal life may assign lower scores to an instructor whose course innovative and challenging just because the student does not want anything that will challenge him/her at school during that term (Rotenberg, 2005).

In addition, halo effect can be considered one of the conditions that threatens the validity of the student evaluations as shown by Mittal, Gera and Batra (2015). These researchers found a halo effect in the use of SETs in their study. They used the two-dimensional scale of SET which was developed by Shevlin et al. (2000) in their study

in the UK and they tested it with Indian students who were pursuing their MBA program in a north Indian university and made necessary modifications. The findings showed that the results of the SET were influenced by the instructor's "Charisma" trait, which was a proof of halo effect (Mittal, Gera, & Batra, 2015). Likewise, Shevlin et al. (2000) found that SETs do not only reflect teaching effectiveness in that how the students perceive the instructor was found to be a significant predictor of student evaluations of teaching effectiveness. This was the charisma factor which explained 69% and 37% of the variation in the "lecturer ability" and "module attributes" factors respectively (Shevlin, Banyard, Davies, & Griffiths, 2000).

Another issue about student evaluations of teaching is the links between the evaluations and student learning. A reanalysis of previous studies by Uttl, White and Gonzalez (2017) found the results of previous studies as inaccurate findings due to their small sample size and the publication bias involved. This current analysis did not show any significant relationship between SET ratings and student learning. Therefore, what the study suggests for the universities is that the universities may want to avoid of SET ratings as a measure of faculty members' teaching effectiveness (Uttl, White, & Gonzalez, 2017).

Similar findings were also found by Stehle, Spinath and Kadmon (2012). Their sample consisted of 883 medical student who were taking one of 32 sections of the same course. Their analysis of the relationship between student evaluations of teaching students' scores in one practical and one multiple choice exam revealed a positive association between SETs and the practical examination where students had to apply the knowledge they gained to the real-life tasks. However, no relation was found between the between SETs and multiple-choice test scores. Therefore, not all tests and all kinds of tests may prove the relationship between student learning and SETs. The authors suggest that some aspects of student learning may be less under the impact of teaching effectiveness (Stehle, Spinath, & Kadmon, 2012).

In addition to these validity issues regarding student evaluations of teaching effectiveness, grade bias is also discussed in the literature. Cho, Baek and Cho (2015) concentrated on the likelihood of an implicit trade-off between students' course grade and SET scores. They introduce the term surplus grade which is the difference between the student's own grade expectation and their actual grade. The results of their study revealed that when this surplus grade is greater than expected, it affected SETs positively. They suggested that the exchange between surplus grade and SETs takes the form of awarding faculty with higher SET scores.

Christiaens et al. (2014) define two proxies of actual learning as "objective measure" which is related to student achievement, and "subjective measure" which is related to students' own perceptions of learning. Their analysis of 1080 SETs delivered in 17 courses in the academic year 2011-2012 showed that both measures of student learning had positive influences on SETs; however, the subjective measure was found to be more powerful. This means that students who think they have learned better give better evaluations to the faculty. This brings the question of objectivity to the SETs.

Nowell (2007) conducted a study to examine the relationship between SET scores and students' relative grades. The data were collected in 32 different courses during the last week of 2003 fall term at a large public university in the United States. Students were asked to give information about their expected grade and were given an instructor evaluation form. A relationship between SET scores and expected grade which was relative to the average grade in an individual class and relative to the average grade given by the instructor. The author claims that they have a powerful proof that students honor instructors not only for their own high grades, but also for high grades received by their peers.

Despite these studies which mention a link between higher grades and higher SET scores, there are also studies whose results deny these findings. Remedios and Lieberman (2008) also questioned if student evaluations of teaching provide valid

measurements of teaching quality. The study was conducted at a Scottish university and all students studying psychology were given two questionnaires, one at the registration and one at the beginning of the following term after they completed the course. Structural modelling of the data displayed that SETs were largely affected by how much they felt involved, how much they found the course "stimulating, interesting and useful", which all implied teaching quality. As to grades and course difficulty, their impact on the end-of-semester ratings was very small.

Moreover, there is vast literature on gender bias of student evaluations of teaching. Basow, Phelan and Capotosto (2006) conducted a study in which 175 students at a large public university described their best and worst professors. The results displayed that male professors constituted 71% of the best professors mentioned by students, and the assets assigned to the best male and female professors again matched gender stereotypes. Best female professors were perceived as more approachable and passionate while best male professors were perceived to be more knowledgeable, passionate and innovative. What this tells us is that if students rate a male professor more knowledgeable in a student rating instrument just because of the male stereotypes in their mind, this requires attention to prevent possible gender bias in the interpretation of the results.

Similar results were also found in another study. Wagner, Rieger and Voorvelt (2016) investigated the impacts of teacher gender and ethnicity on student evaluations of teaching at university. The study showed an adverse impact of being a female teacher on student evaluations of teaching. It suggests that there is a gender bias to the detriment of female instructors in student evaluations of teaching and therefore attention should be paid to these issues while using student evaluations for the purposes of hiring and promotion (Wagner, Rieger, & Voorvelt, 2016).

On the other hand, there are also other studies which found that gender bias in student evaluations of teaching are not significant enough to make these evaluations biased. Punyanunt-Carter and Carter (2015) gave 58 students in a basic introductory

communication course an altered instructor evaluation form. Half of the students answered the form for a male professor, and the other half answered the questions for a female professor. Punyanunt-Carter and Carter argue that although the results showed some gender bias, the bias was not statistically signicant to have impact on the evaluation results.

Another study which did not find large gender differences between male and female instructors was conducted by Centra and Gaubatz (2000). They conducted the research in 741 classes which had at least 10 male and 10 female students. In two different analysis, student ratings for male and female professors; and student ratings by male and female students were compared. The results showed same gender preferences especially in female student female professor ratings. However, Centra and Gaubatz claim that these differences are not large and not expected to cause differences in personnel decision. They also discuss if gender differences are because of the differences in teaching styles.

Although there are not conclusive results in the literature about gender, grade knowledge and other bias in student evaluations of teaching, they continue to be used as a form of faculty evaluation. AL-Saghir (2008) states that although in the literature there are both implicit and explicit criticisms regarding the utility, reliability and validity of SETs against various aspects of effective teaching and potential biases, the research continue to indicate that SET scores can be useful for the faculty members, the administration and the students (AL-Saghir, 2008). However, for these evaluations to be useful, the administrators still need to be aware of these possible biases if they want to use the results for summative purposes. Otherwise, when the gendered nature of universities and their gendered practices are considered, these possible biases may affect the results and destroy the validity of the instruments.

2.2 Gender in Job Performance Evaluation

2.2.1 Gender and gender stereotypes in organizations

Now that teaching effectiveness and student evaluations of teaching have been discussed, it is critical to take a look at gender issues in performance evaluation and teaching effectiveness in higher education. West and Zimmerman (1990) define gender as an attained state whose formation is subject to psychological, cultural and social processes. In their definition, gender is something that is "done", creating unnatural, inessential and artificial differences between men and women. However, the problem with this artificial construction is that once the differences are built, they serve the purpose of strengthening the importance of gender. Through appropriation of gender ideals and identities, these essential gender differences, which were once unnatural and inessential, turn into "objective facts". In this way gender differences become normal and natural; and consequently, legitimize the different male and female destinies in the society (West & Zimmerman, 1990).

In this legitimization process, Lorber (1994) reminds us that all society, its values and institutions like religion, science, and law play a role. Lorber also adds that these legitimized differences seriously restrict or provide opportunities for male or female accomplishments while also affecting the quality of these accomplishments. In sum, then, gender becomes a strong ideological means that plays a role in the production, reproduction, and legitimization of individuals' options and boundaries which are based on sex category (West & Zimmerman, 1990). As we have seen although gender differences are not necessarily true differences between men and women, they join our lives as normal and legitimate. The same is valid for gender stereotypes. By definition, a stereotype is an unjust and unreal notion that people hold for all individuals or things that share a specific characteristic or feature (Online Merriam-Webster Learner's Dictionary, n.d.). As to the gender stereotypes, Ashmore and Del Boca (1979) define them as structured assumptions which connect individual characteristics to the societal categories of women and men.

Basow (1992) acknowledges that gender stereotypes have a strong effect as a means of social control although only few of them address true differences between men and women. Basow claims that they create self-fulfilling prophecies. Passer and Smith (2008) maintain that self-fulfilling prophecy happens when individual's behaviors are shaped by fallacious expectations from others, which, in turn, causes the expected behaviors to come true, therefore verifying their first impression. In other words, groundless gender-based assumptions we carry influence our acts toward other people, modifying their behaviors in a way that verify our expectations from them. Reeves (2010) exemplifies these assumptions by elaborating on a manager who has an untrue anticipation of a female employee. Although the expectation here is untrue, the manager's acts will have an impact on the employee and the employee will start to act in ways that will match these untrue expectations. As a result, if untrue expectation soars the view that female employee is incapable of doing something, the chances that she can manage it are reduced by this expectation.

When Basow's warning about gender stereotypes and self-fulfilling prophecies are considered, it become critical to how different genders are depicted in the workplace. By investigating 219 discrimination stories drawn from sex discrimination cases worked upon by the Ohio Civil Rights Commission, Bobbitt-Zeher (2011) concluded that women faced many stereotypical assumptions in the workplace. The women workers were firstly deemed as women, and then as worker. The way their body looked and the way they behaved were expected to be in harmony with the assumed gender stereotypes for women, and otherwise they were punished for their "unladylike" behaviors. Although men were not punished for the use of inappropriate words and aggression in the workplace, the opposite case was true for women. In the study, gender stereotypes were also found in the way women were depicted as unclever, hormonal and too sentimental (Bobbitt-Zeher, 2011).

Universities are also, as a workplace, are not free from being gendered. Bird (2011) agrees that universities are also gendered bureaucratic organizations and suggests that although there are systematic barriers which need to be removed to help women

progress in their careers, university administrations and faculty members have not been quick to notice these barriers. One reason for this situation stems from the fact that people at the superior decision-making positions continue to produce women-centered explanations for women's absence in these ranks (Bird, 2011).

Similarly, assuming that gender is socially constructed as a result of gendered "performances", Lester (2008) set off from performativity framework and interviewed with female faculty members to understand the production and reproduction of gender roles at a community college. The results of her study showed that credibility of female faculty members was frequently challenged by both students and colleagues. One of the participants who was working as a welding instructor in a male-dominated department reported that she felt a need to behave toughly and sternly in order to retain respect from students and to suppress the challenges to her credibility although she was not that kind of person out of the school; and this caused her to disentangle her gender performance at work and her female identity. However, although they had to act more toughly and sternly to gain respect from students, female faculty was also required to play their mother roles by sustaining an expressive and caring classroom environment (Lester, 2008).

The discussion here about gender and higher education institutions present us two conclusions. One of them is that workplace is gendered, which makes higher education institutions and their practices also gendered. The second conclusion is that people are expected to behave according to the expected gender stereotypes in the workplace although these gendered assumptions are not always true.

2.2.2 Gendered nature of job performance evaluation

If workplace and its practices are gendered, job performance evaluation also need to be handled with this perspective. To this end, it is important to start with the definitions of "job" and "job evaluation" to understand the gendered nature of job evaluation processes in organizations. Acker (1990) defines the job as a key

component in the hierarchical structure of a work organization and regards it as a portrayal of a group of chores, capabilities, and duties which constitute a position in the organizational chart. Acker describes job performance as a process in which the content of jobs is defined and the comparisons among jobs are done based on expertise, abilities, difficulty, required effort and conditions of employment. On the basis of these two definitions, Acker problematizes the logic behind job evaluation whose aim is to evaluate the characteristics of the job, not of the people who perform the job and the fact that a job is independent of people who perform it. According to Acker, the point which requires attention in these definitions is the expectation of a hypothetical worker who does not have a body on the grounds that workers have gendered bodies in actual situations. Therefore, Acker maintains that this abstract, disembodied worker who holds a gender-neutral position in the organization is not realistic and serves to mask and recreate undisclosed gender relations in organizations. This view is also supported by Ridgeway (2009). Ridgeway (2009) claims that gender is embedded in organizations and understanding the interaction between the organizational rationale and the underlying effects of gender frame is a key to grasp the outcomes of certain organizational processes and the gender structure they create (Ridgeway, 2009).

Bailyn (2003) criticizes the definition of gender equity which is regarded as gender equality. By dictionary definition, while equality is "the state of being equal", equity involves "fairness or justice in the way people are treated" (Merriam-Webster Learner's Dictionary, n.d.). Therefore, according to Bailyn, the practice of presenting equity as equality comes with the assumption that workplace is distinct from the other aspects of life and disregards the lives which take place out of the workplace. This definition cannot grasp the differences between the life experiences of men and women outside the academic work and sets the ideal academic worker norm as "male" (Bailyn, 2003).

Burton (1991) argues that one of the drawbacks of job evaluation systems is their potential contribution to the institutionalization of gender bias in organizations and

presents a critique of job evaluation related to the data collection procedures. The data in these evaluations can be collected through interviews, questionnaires, reports and similar techniques; however, it should be kept in mind that way the statements are verbalized and the way the instruments are designed may have effects on the data and results of the evaluation. The bias may either stem from people's gender-related perceptions or blindness to the requirements or characteristics of female jobs (Burton, 1991).

Gender bias in performance evaluation processes has also been supported by other studies in the literature. Chung, Marshal and Gordon (2001) found a gender bias in supervisory evaluation. The participants of the study were given 4 hypothetical supervisee profiles, and these profiles were the same except the supervisee being Black male, White male, Black female or White female in each case. Male participants gave more negative ratings when the supervisee profile was presented as female. However, female participants did not show such gender bias in their responses (Chung, Marshall, & Gordon, 2001).

Furthermore, in a study whose aim was to understand how men and women in engineering and science teams evaluate their colleagues' expertise, Joshi (2014) found that highly educated female participants received more negative evaluations when compared to their less-educated colleagues by male participants who firmly associated themselves with their gender. However, the same pattern was not valid for female participants. Female participants made more positive evaluations for highly educated male and female colleagues (Joshi, 2014).

In a study to investigate the role of gender in self and supervisory performance evaluations, Wren (2006) found that women perform as well as men in all skills except task-specific skills in which they gave themselves a lower score. There was not significant pro-male gender bias in supervisory evaluations, and women were rated higher in establishing social relationships with their customers, colleagues and supervisors. When compared to the supervisory ratings, women overrated their skills

only in one dimension, while men overrated their skills in three of four dimensions stated in the performance evaluation criteria (Wren, 2006).

To summarize briefly, what can be concluded from the discussions above is that higher education institutions are not gender-free organizations, and therefore jobs and job evaluation systems that are conducted in these organizations should also be investigated through a gender perspective. Otherwise, there is a threat for gender equality in the workplace. The discussions above show us that this threat can be strengthened in different ways.

Firstly, the threat for gender equality is strengthened through generation of women-centred explanations for the inequality as Sandberg (2013) claims. She maintains that individuals' own preferences are not as personal as they may seem and shaped by social and familial expectations and pressures. Nussbaum (2000) also rejects "utilitarian preference-based approaches". Accordingly, customs, apprehensiveness, low-set expectations and unfair personal backgrounds ruin individuals' preferences and aspirations for their own lives. Therefore, putting the blame on women and claiming that "it is their own choice not to become school leaders" only masks the problem. What we need to do to bring equality is to remove systematic barriers which work against women as Bird (2011) suggests.

Secondly, our neglect of ideal academic worker who is set as a man as claimed by Bailyn (2003) serves to strengthen gender inequalities in higher education institutions like other workplaces. As put forward by Acker, this academic ideal worker also does not have any familial responsibilities out of the workplace, and body and behaviour expectations from this ideal fit more in male stereotypes. Lester's 2008 study, as mentioned earlier, is a good example of how gendered expectations can shape life and work experiences of a female faculty member. To fulfil this male worker ideal, while sustaining an expressive and caring classroom environment, she feels a need to behave toughly and sternly in order to retain respect from students and

to suppress the challenges to her credibility although she was not that kind of person out of the school.

What is common in these two explanations is, indeed, the neglect of higher education institutions as gendered workplaces as well as social and structural inequalities. These social and structural inequalities show themselves in leadership positions or in job performance evaluation systems used in these institutions. Even when it is easy for someone to claim that there is no gender bias in performance evaluation process, as Burton's (1991) study shows that there may be some bias in the way the data is collected; through interviews, questionnaires, reports and similar techniques; or in the way the statements are verbalized, or the way the instruments are designed. All these can have effects on the data and results of the evaluation. According to Burton, also, these can originate from either people's gender-related perceptions or blindness to the requirements or characteristics of female jobs.

If there is a threat that job performance evaluation instruments can carry gender bias due to taken for granted gender-related perceptions, it is important to examine student evaluations of teaching effectiveness with this perspective. For example, in the teaching effectiveness instrument used in this study, there are items like "The instructor was knowledgeable in his/her field", or "The instructor was warm and friendly" which may carry some gender bias in that there is research which showed that students have a tendency to regard male professors more knowledgeable (Boring, 2016) and they reported a higher level of rapport as to female professors (Joye & Wilson, 2015).

This type of items may prove Burton's argument that the way statements are verbalized in these instruments or the way they are designed may be gender-biased. As mentioned earlier, therefore, it is important to question student evaluations of teaching with a gender perspective. Although the literature presents facts against this argument, there is also a huge amount of literature which claims that there is gender bias in student evaluations of teaching.

2.2.2.1 Gender and student evaluations of teaching effectiveness

The studies presented in this part displays that gender stereotypes and distinct aspects of teaching are not independent of each other. MacNell, Driscoll and Hunt (2014) investigated gender bias in student evaluations of professors in an online course by misstating the gender of the assistant instructors. The instructors who were presented as female identity got lower scores irrespective of their real gender. In addition, the same acts of professionalism/unprofessionalism were considered differently by students depending on professor's perceived gender. To exemplify, when in the case where actual male and female instructors who were falsified as male sent grades two days later, this act got 4.35 out of 5 in terms of promptness. However, when the same instructors were presented as female and sent grades two days later, this act got 3.55 out of 5 in terms of promptness. The authors argued that being a male instructor leads to an automatic recognition of reliability as to their professionalism, expertise and effectiveness (MacNell, Driscoll, & Hunt, 2014). This study and similar studies mentioned below clearly show us that gender stereotypes that students hold may have impacts on student evaluations of teaching effectiveness.

Similarly, Miller and Chamberlin (2000) did a study to investigate if students hold different perceptions of educational credentials of male and female instructors. The findings revealed that male instructors were thought to hold higher educational attainments when compared to women. For example, a Ph.D. attainment was seen as more likely to be held by a male graduate instructor in comparison to a female full professor.

Baker and Copp (1997) examined the changes in students' evaluations in relation to the degree that the professor could meet students' gendered expectations. They used both qualitative and quantitative teaching evaluation data from Dr Baker's feminist course, three terms of which she experienced pregnancy. The results indicated that students reacted differently to the professor when they were in a small summer class where the professor could satisfy gendered expectations, students were more

welcoming of a feminist perspective and their professor's pregnancy; and when they were in a large fall class where the professor could not fully introduce herself as attentive, warm and amiable (Baker & Copp, 1997).

Takiff, Sanchez and Stewart (2001) did two studies to examine the differences in students' ways of addressing male and female professors. The results of their first study demonstrated that students were likely to address male professors by titles when compared to female professors. The results of the second study, in addition, disclosed the fact that a higher status was attributed to professors when they were addressed by their titles; however, when female professors were addressed by their titles, they were regarded as less reachable. These two studies showed that female professors were perceived to have lower status in comparison to male professors and prestige and approachability are usually two mutually exclusive terms for them (Takiff, Sanchez, & Stewart, 2001).

Joye and Wilson (2015) studied the impacts of professor gender and perceived age on student evaluations of effectiveness, rapport and academic performance. The results exhibited a higher level of rapport and attractiveness as to female professors, and a higher level of effectiveness as to male professors. Also, younger female professors were perceived to be more attractive when their scores are compared to other conditions in the study, which meant that gender and age intersect in student evaluations of their professors.

Carli et al. (2016) undertook two studies to analyze the commonalities between gender stereotypes and successful scientist stereotypes. The results of the first study whose data were drawn from 180 female undergraduate students at a liberal arts single-sex college and 73 female and 61 male undergraduate students at a private university manifested that while there were marked similarities between the male stereotypes and successful scientist stereotypes, these similarities were not valid for female stereotypes and successful scientist stereotypes, and this was found to be true for all three groups in the study: female undergraduates in the single-sex college,

male students in the mixed-gender university and female students in the mixed-gender university. However, one important result of the study was that female students at single-sex college more powerfully made connections between the characteristics of successful scientists and women. Their second study, whose participants were 294 female undergraduates at a liberal arts single-sex college, and 225 female and 116 male students at co-educational universities and colleges in the United States, also supported the findings of the first study and revealed stronger similarities between male stereotypes and successful scientist stereotypes especially in the fields of biology, chemistry, physics and computer science. Nonetheless, the similarities between female and successful scientist stereotypes were stronger in the field of psychology. They concluded that as the number of women in a specific field increases, the characteristics of women and scientists get closer to each other (Carli, Alawa, Lee, Zhao, & Kim, 2016).

All the studies mentioned above point to possible relationship between gender of the faculty member and his/her perceived teaching effectiveness. However, there are also studies in the literature which did not find significant impacts of gender on student evaluations. Tindall and Waters (2017) studied student evaluations of public relations educators, and they found that female professors were regarded as showing more professional warmth, more professional competence and being more connected to public relations industry while male professors were regarded as delivering more difficult courses although statistical significance levels have not been met (Tindall & Waters, 2017).

By reviewing the literature on gender and student evaluations of teaching, Feldman (1993) found that although the average correlation favored women with higher evaluation scores, this correlation was too small and unimportant to affect the results even when it was statistically significant. Feldman also underlined the inconsistencies in the findings of the existing research by highlighting the fact that under different conditions gender-typical and gender-atypical behaviors and characteristics led to higher evaluation scores. In addition, the interaction effects of gender with academic

title, discipline, course level, pedagogical beliefs of the instructor and personality were also inconsistent across studies. As an example of gender atypical behaviors leading to positive student evaluations, a study by Meltzer and McNulty (2011) can be discussed. In a study which asked participants to evaluate hypothetical job candidates, Meltzer and McNulty found that the unexpected characteristics of a sex can turn into an advantage in professor evaluations. Participants in the study favored the "nurturing" male professor more than the corresponding female professor (Meltzer & McNulty, 2011). Another example of gender atypical behavior turning into a benefit comes from Yamawaki et al. (2012). They created hypothetical scenarios to study the impacts of an instructor's sex, field of study, gendered teaching style and expected grade on teaching evaluations. The field of study and instructor's sex interacted significantly, and the results showed that a female professor in the department of computer science was considered to be more knowledgeable when compared to a male professor in the same department. Likewise, a male professor in the department of psychology was considered to be more knowledgeable in comparison to a female instructor in the same field. It was concluded that the professors of opposite gender in stereotypically male or female fields are perceived to have more ability and knowledge.

Wheeless and Potorti (1989) investigated if gender differences had an effect on student attitudes as to learning. The research examined the interaction between gender and gender role characteristics and the results indicated that this interaction was not significant to affect student learning. Instead, student learning was found to be in correlation to masculinity or femininity of the instructor, irrespective of whether the instructor was male or female. Similarly, gender of the student was not found to statistically significant in the analysis. The overall results show that the personality characteristics of the instructor, dominant/masculine and expressive/feminine, was found to be more important than whether the instructor or the student was male or female.

Fernángez and Mateo (1997) investigated if the interaction between student gender and faculty gender had a significant impact on student evaluations of teaching taking different academic fields as a modulating variable. A sample of 1,304 student were selected from a university in Madrid, Spain; and the results did not show any statistically significant interaction between student gender and faculty gender to have an impact on student evaluations of teaching.

Price et al. (2017) investigated student evaluations of teaching quality in courses in the programs computer science and environmental engineering at a large Swedish university. The data consisted of 8888 sets of ratings for men and 4280 sets of ratings for women from the Course Experience Questionnaire over ten academic years. The results showed some differences for teachers with higher scores in courses which are less typical for their gender. However, when the overall results are considered, the difference between student gender and the differences between faculty gender revealed only small effects, which had only little theoretical or practical significance. As can be seen from the literature above, there are inconsistent results about the impacts of gender on perceived teaching effectiveness of faculty members. However, when all studies are taken into consideration, studying these gender biases becomes vital. Students can have gendered expectations from their professors regarding their professionalism, prestige, credibility, approachability, effectiveness and attractiveness, and this can turn into an advantage or disadvantage for one gender depending on what is expected. As student evaluations of teaching are used for both summative and formative faculty performance evaluation technique, the administrators of the instrument should also keep in mind these possible bias in the analysis and interpretation of results. It is also vital to pay attention to researchers like Wagner, Rieger, and Voorvelt (2016) who suggested that there is a gender bias in teaching evaluations which place women in a more disadvantageous position in hiring and promotion decisions. When this bias is ignored, gendered nature of higher education institutions will be reproduced through such practices.

2.3 Summary

The purpose of this study was to investigate if there are any statistically significant differences in student evaluations of teaching effectiveness scores by faculty members' gender and student characteristics which are students' gender, course grade and discipline. This chapter presented existing literature regarding the research purpose. The first section presented a discussion of faculty performance evaluation in higher education. Faculty performance evaluation is defined as an ongoing and standardized process (Arnautu & Panc, 2015) and it is used for both summative and formative evalution purposes in higher education (Channing, 2016). However, as its uses are not free from problems, there is a need for well-functioning evaluation systems in higher education (Kalaycı, 2009). Therefore, three different faculty performance evaluation techniques were discussed including their advantages and disadvantages. One of these techniques was student evaluations of teaching effectiveness, which was questioned in the existing literature mainly because of the changing definitions of teaching and teaching effectiveness. As there is not only one accepted definition of these concepts (Ryans, 1949; Arreola, 2007), some part of literature questions the validity of these instruments (Spooren, Brockx and Mortelmans, 2013). Other validity issues regarding student evaluations of teaching effectiveness also include possible gender bias (Das and Das, 2001) and students' grade knowledge (Cho, Baek, & Cho, 2015) among other validity issues.

The second part of the literature review focuses on gender in job performance evaluation. Since the concepts of teaching effectiveness and student evaluations of teaching effectiveness have been discussed in the previous part, it is critical to look at gender issue in these evaluations which are a form of faculty performance evaluation in higher education. The main arguments in this part are, firstly, that universities as a workplace are gendered institutions, which also makes their practices gendered (Bird, 2011); secondly, that people are expected to behave according to gender stereotypes in the universities (Lester, 2008); and thirdly, job performance evaluation as its conduct is not free from gendered expectations, is also gendered

(Acker, 1990). Therefore, student evaluations of teaching effectiveness as a form of faculty evaluation require attention from a gender perspective.

Although there is research on gender issues in student evaluations of teaching effectiveness which indicated no gender bias (Punyanunt-Carter and Carter, 2015), there is also research which indicated significant gender bias in student evaluations of teaching effectiveness (Boring, 2016). When all these studies and discussions are considered, it becomes vital to investigate if there are any statistically significant differences in student evaluations of teaching effectiveness scores by faculty members' gender and student characteristics. This chapter presented the related discussions and studies regarding the research questions of the present study. The next chapter, which is Methodology, presents the research design used.

CHAPTER 3

METHODOLOGY

This chapter presents the methodological procedures followed in this study. The design of the study, population and sample, variables of the study, data collection procedure, ethical permission, data analysis, internal and external validity and limitations of the study are covered as the major topics of this chapter.

Quantitative research with a causal comparative design was employed as the research design of this study. The reason for the selection of this method is that it investigates the differences in student evaluations of male and female faculty members, and the present study does not allow for any manipulation or treatment. The target population of the study is university students and the sample come from students at Middle East Technical University.

3.1 Design of the Study

Causal comparative design was used for this quantitative research. Fraenkel, Wallen and Hyun (2012) claim that studies with a causal comparative design intend to discover the cause or consequences of already-existing differences among individuals or groups. The variable which is different between/among groups is not manipulated in this type of studies and at least one variable is categorical (Fraenkel, Wallen, & Hyun, 2012). Accordingly, this study aims to reveal already existing gender differences, grade-based and discipline-based differences in student evaluations of faculty members' teaching effectiveness.

3.2 Population and Sample

All university students in Turkey constituted the population of the study. The study was conducted in the capital city of Turkey, Ankara. The target population of this study was university students. The sample was selected from Middle East Technical University, and students at Middle East Technical University were asked to participate in the study for convenience reasons. The sample consisted of 667 students from 27 different departments at Middle East Technical University. Among these, 256 students were in their sophomore year, 191 students were in their junior year, and 220 students were in their senior year at university.

3.3 Variables of the Study

Research question 1: Is there a statistically significant difference in student evaluations of teaching effectiveness scores by faculty member's gender and student's gender?

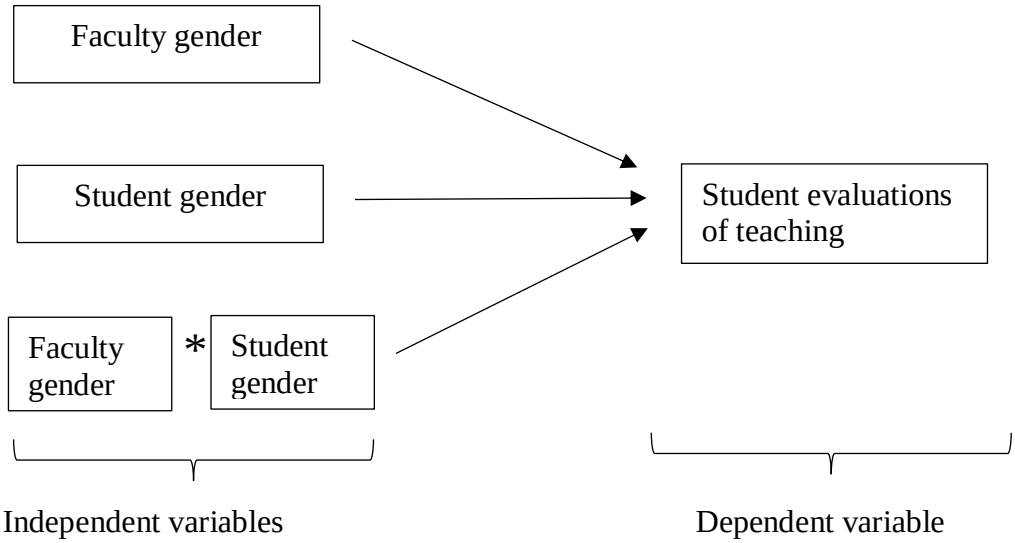

Figure 3.1. Demonstration of variables of the study for the first research question

Research question 2: Is there a statistically significant difference in student evaluations of teaching effectiveness by student's grade?

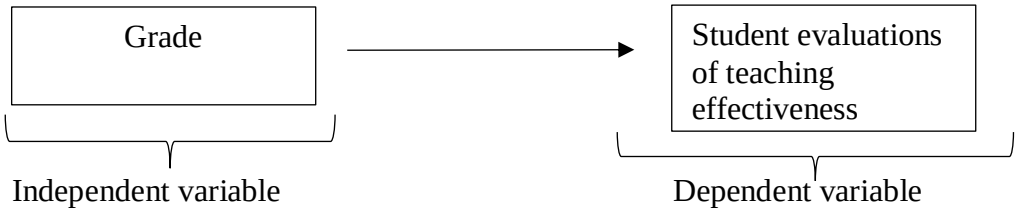

Figure 3.2. Demonstration of variables of the study for the second research question

Research question 3: Is there a statistically significant difference in student evaluations of teaching effectiveness by student's discipline?

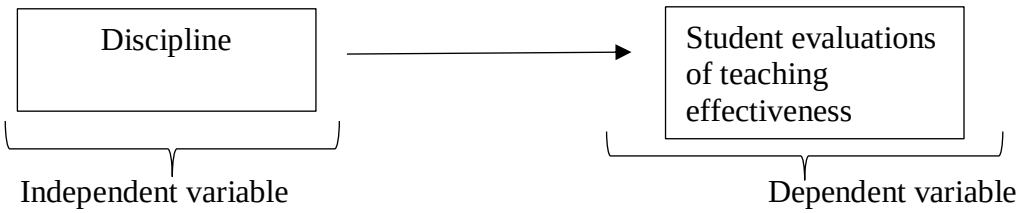

Figure 3.3. Demonstration of variables of the study for the third research question

3.4 Data Collection Instruments

Data were collected with an instrument which had two parts: Demographic information and instructor rating questionnaire. The instructor rating questionnaire was a pre-developed questionnaire. The following section delivers comprehensive information about the data collection instruments used in the study.

3.4.1 Demographic information

This part of the instrument asked six questions about the characteristics of the participants. The categorical variables investigated are what year of college education the student is in, student's department, the selected faculty member's gender, student's gender, student's cumulative grade point average (CGPA), and student's final grade in the selected faculty member's course.

3.4.2 Instructor rating questionnaire

To analyse teacher effectiveness, Instructor Rating Questionnaire, which was developed by Young and Shaw (1999), was used. The questionnaire consists of 25 items, and it aims to measure teacher effectiveness in terms of value of the course, motivating students, course organization, effective communication, concern for student learning and genuine respect for students. The adaptation of the questionnaire into Turkish was done by Emil (2013).

This part of the instrument required participants to select one of the instructors whose course they have taken in their university education as the most/least effective instructor and to rate how effective their teaching was by answering 25 items in the questionnaire.

Sample items from the scale are "Öğretim elemanı kendi alanı hakkında bilgiliydi. / The instructor was knowledgeable about subject matter.", "Öğretim elemanı sıcak ve arkadaşça davranırdı. / The instructor was warm and friendly.", and "Öğretim elemanına ders dışında da ulaşılabilirdi. / The instructor was accessible outside of the class.".

3.5 Data Collection Procedure

The data collection procedure lasted for a month in the Spring semester of 2017-2018 academic year. The data were collected from 667 students through random visits to classrooms. Before these visits, the professors were informed about the data collection procedure and their permission to enter classrooms was taken. Once the permission was taken, the time of the visit was arranged. Then, the surveys were distributed to students on the basis of volunteering after they had been informed about the data collection procedure and ethical permission of the study. The procedure followed these steps: Students were asked to choose the faculty member they thought as either the most effective or ineffective, the course of the faculty member chosen

should be completed, and students were asked to share neither their own identities nor the identity of the faculty member they had chosen. The procedure was completed in approximately 15 minutes in each classroom visited.

3.6 Ethical Permission

Prior to data collection, the ethical permission from the Middle East Technical University Human Subjects Ethics Committee (Appendix A) was obtained. The data collection instruments used in the study do not require students to state their names, and in this way the confidentiality of the data was secured. Participants of the study were provided with the information that they have the right to leave the study any time they wished.

3.7 Data Analysis

In this study descriptive statistics and inferential statistics are employed for data analysis and the interpretation of results through the use of Statistical Package for Social Sciences (SPSS) 23 program. Data analysis was conducted to investigate whether there are statistically significant differences between group or the differences exist due to a coincidence. For this purpose, descriptive statistics: mean (*M*), median, standard deviation (*SD*), skewness and kurtosis were calculated; the assumptions of two-way ANOVA were checked; and inferential statistics: two-way ANOVA was run for the first research question.

The reason to use two-way ANOVA to investigate the differences between student evaluations of teaching depending on student's gender and faculty member's gender was the presence of two independent categorical variables which are student gender and faculty gender in the research structure (Gravetter and Wallnau, 2013). Two-way ANOVA allowed us to see if there are statistically significant differences in student evaluations of teaching effectiveness scores when student gender, faculty gender and the interaction between them are taken into consideration.

For the second and third research questions, descriptive statistics: mean (*M*), median, standard deviation (*SD*), skewness and kurtosis were calculated; the assumptions of one-way ANOVA were checked; and inferential statistics: one-way ANOVA was planned for each question. However, for the two independent variables of these analyses as the results revealed a violation of Levene's test of homogeneity of variances, Welch's F test and its follow-up analyses were conducted for the independent variables grade and discipline for the second and third research questions.

Field (2018) argues that when homogeneity of variances assumption is violated, either Brown-Forsythe or Welch's test can be used as a correction to this violation. As Welch's F test is better at revealing an existing effect (Field, 2018), it was preferred for this study. The Welch's F test allowed us to if there are differences between student evaluations of teaching effectiveness scores firstly by focusing on student's final course grade and then by focusing on the discipline they have been studying in.

3.8 Internal Validity

Among the sources of invalidity associated with causal comparative design are subject characteristics, instrumentation, and they need to be paid attention in this study.

By definition, an important subject characteristic included at which year the participants of the research study. To solve the first internal validity problem and to ensure that they have met enough number of professors to make comparisons among them, 2nd, 3rd and 4th year students are chosen to participate in the study.

As to the instrumentation, it is another internal validity threat. By definition, instrument decay takes place when the instrument is too long or difficult to score. The scale consists of 25 items, so this internal validity problem is overcome.

3.9 External Validity

The target population in this study were college students in Turkey. Students at Middle East Technical University were selected as the sample for convenience reasons. Random sampling method could not be employed in this study because the data were collected in the classrooms and the permission of the instructors to be able to collect data was necessary. The sample consisted of 667 students, however, the generalizability of results to all college students in Turkey can be contestable in that random sampling could not be used. The threat for external validity is that not all individuals in the population had equal chances to participate in the study.

3.10 Limitations of the Study

The present study had a causal comparative design, so the data was collected through a questionnaire and the analysis of the data does not reveal information regarding the causes of the phenomenon. To have a deeper understanding of the reasons for students' evaluations of faculty member's teaching and the role of faculty member's gender in this evaluation, it would be beneficial to hold interviews with students.

The study was conducted at Middle East Technical University, and the participants of the study were undergraduate students in different departments. In addition, due to English being medium of the instruction and the international campus culture, the students' perceptions of gender stereotypes may not represent the perceptions of the population.

Also, the intersection of faculty member's gender, age, title and managerial roles together could be an issue in how students perceive the faculty member's teaching effectiveness. However, this study did not focus on such intersections and remains a descriptive study presenting only the differences in the scores of student evaluations of teaching effectiveness when student gender and faculty member's gender are taken into consideration

CHAPTER 4

RESULTS

The findings of the study are given in this section. Firstly, descriptive statistics are presented, secondly the assumptions of the statistical test that is used are checked and presented, and lastly the findings of the study are given.

4.1. Descriptive Statistics

In this part, descriptive statistics are shown to present a general picture of the data with the aim of making it easier to understand. Descriptive statistics of the data set was run by SPSS 23, and they include the number of participants (N), minimum and maximum, mean (M) and standard deviation (SD).

A total number of 667 students participated in the study. 390 of the participants were female (58.5%) and 276 of the participants were male (41.4%), and 1 student (0.1%) did not identify their gender. Among these 667 participants, 256 of them were in their sophomore year (38.4%), 191 of the participants were in their junior year (28.6%) and lastly 220 participants were in their senior year (33%) at the university. Students from 27 different departments were represented in the study and the distribution according to their disciplines was as following: 75 of the participants (11.3%) were students of Architecture, 63 of the participants were students of Humanities (9.4%), 216 of the participants were students of Engineering (32.4%), 104 of the participants were students of Natural Sciences (15.6%), 93 of the participants were students of Economics and Administrative Sciences (13.9%), and lastly 115 of the participants were students of Educational Sciences (17.2%).

Since the main purpose of the study was to investigate whether male and female faculty member's student evaluation scores differ, students were asked to choose a faculty member whom they find either effective or ineffective to be evaluated and indicate their gender. According to the descriptive statistics, 294 female faculty members (44.1%) and 373 male faculty members (55.9%) were selected by the participants of the study to be evaluated in terms of their teaching effectiveness.

Table 4.1.

Summary Table for Participant Characteristics

Participants	N	% Percentage
Gender		
Female	390	58.5
Male	276	41.4
Other or not defined	1	0.1
Year at university		
Sophomore	256	38.4
Junior	191	28.6
Senior	220	33
Discipline		
Architecture	75	11.3
Humanities	63	9.4
Engineering	216	32.4
Natural Sciences	104	15.6
Eco. & Adm. Sciences	93	13.9
Educational Sciences	115	17.2

As to the question 26 of the instrument which is about teaching effectiveness of the selected faculty member as effective or ineffective when compared to the other faculty members whose courses were taken, Table 4.2. displays that most of the students chose the faculty members whom they find effective.

Table 4.2.

Descriptive Statistics for Teaching Effectiveness of Selected Faculty Member (Q 26)

	N	% Percentage
Not applicable	2	.3%
Strongly disagree	160	24%
Disagree	77	11.5%
Undecided	39	5.8%
Agree	78	11.7%
Strongly agree	311	46.6%
Total	667	100%

As can be seen from Table 4.3. the minimum total score for the Teaching Effectiveness Instrument was 25 and the maximum total score was 125, the mean score students gave to the selected faculty members was 90.15.

Table 4.3.

Descriptive Statistics for the Total Score of Teaching Effectiveness Instrument

	N	Min.	Max.	Mean	*SD*
Total	667	25.00	125.00	90.15	30.32

In addition, regarding the first research question which is "Is there a statistically significant difference in student evaluations of teaching effectiveness scores by faculty member's gender and student's gender?", total scores in the teaching effectiveness instrument for male and female faculty members were calculated to see the differences between them as well as the differences in the scores given by male and female students to the selected faculty members. Data revealed that female faculty members are evaluated with higher score as to student evaluations of teaching performance (*M*=93.26, *SD*=29.89) when compared to male faculty members (*M*=87.87, *SD*=30.32). Also, female students are found to be giving higher evaluations as to student evaluations of teaching effectiveness (*M*=92.42, *SD*=30.79) when compared to male students (*M*=87.17, *SD*=29.20).

Table 4.4.

Descriptive Statistics for the Male and Female Faculty Members' Teaching Effectiveness Scores

	N	Mean	*SD*
Female	294	93.26	29.89
Male	372	87.87	30.32

Table 4.5.

Descriptive Statistics for the Teaching Effectiveness Scores by Male and Female Students

	N	Mean	*SD*
Female students	390	92.42	30.79
Male students	276	87.17	29.20

Table 4.6. shows that female students gave both higher evaluations to both female faculty members and male faculty members compared to male students. Female faculty members were also found to be receiving higher evaluations both from male and female students.

Table 4.6.

Descriptive Statistics for the Teaching Effectiveness Scores by Male and Female Students

	Mean	*SD*
Female students		
Female faculty	95.27	30.37
Male faculty	89.49	31.04
Male students		
Female faculty	89.10	28.60
Male faculty	86.15	29.54

Regarding the second research question which is "Is there a statistically significant difference in student evaluations of teaching effectiveness by student's grade?", the descriptive statistics given in Table 4.7. showed that in 12 groups, students who got the S grade which is given in non-credit courses gave the highest evaluations as to teaching effectiveness (*M*=116.75, *SD*=6.34), and this was followed by students with the highest grade AA (*M*=107.10, *SD*=23.28). Students who withdrew from the course gave the lowest evaluations (*M*=52.25, *SD*=5.91), and this was preceded by students who got a failing grade which is FF (*M*=54.42, *SD*=13.75).

Table 4.7.

Means and Standard Deviations for 12 Grade Groups

Grades	*M*	*SD*	*n*
AA	107.10	23.28	195
BA	104.68	23.58	98
BB	90.32	27.83	96
CB	83.80	26.16	68
CC	73.36	29.98	65
DC	72.56	28.10	37
DD	70.28	27.57	37
FD	55.86	16.08	30
FF	54.42	13.75	19
S	116.75	6.34	4
W	52.25	5.91	4
NA	60.86	27.55	7
Total	89.97	30.31	660

Lastly, as to the third research question which is "Is there a statistically significant difference in student evaluations of teaching effectiveness by student's discipline?", the descriptive statistics given in Table 4.8. showed that students of Humanities evaluated teaching effectiveness highest (*M*=100.76, *SD*=27.17) while students of Natural Sciences evaluated teaching effectiveness lowest (*M*=82.56, *SD*=30.99).

Table 4.8.

Means and Standard Deviations for 6 Disciplines

Faculty	*M*	*SD*	*n*
Architecture	82.84	29.05	75
Humanities	100.76	27.17	63
Engineering	90.07	30.46	216
Natural Sciences	82.56	30.99	104
Eco. & Administrative Sciences	95.27	28.5	93
Educational Sciences	92.01	31.17	115

4.2 Inferential Statistics

Inferential statistics were used to analyze the data to have a more detailed picture of what descriptive statistics mean. In this part, inferential statistics for each research question were presented.

4.2.1 Inferential statistics for research question 1

The first research question was "Is there a statistically significant difference in student evaluations of teaching effectiveness scores by faculty member's gender and student's gender?" As the question has two independent variables which are student gender and faculty member gender and one dependent variable which is the score on the Instructor Rating Questionnaire, two-way ANOVA was conducted to see if there is any statistically significant difference between male and female faculty members teaching effectiveness scores depending on these two independent variables and the interaction between them.

4.2.1.1 Assumptions of two-way ANOVA

Before we go through the two-way ANOVA statistics, we need to check if assumptions of test have been satisfied. Firstly, the data were gathered from

participants in the classrooms that they were taught and therefore students may have interacted with each other and chosen the same faculty members during the implementation. This situation means that the independent observations assumption is violated. Secondly, normality was checked for both faculty member's gender and student's gender and the interaction between faculty member's gender and student's gender through 4 different means: Skewness-Kurtosis, Normality tests (Kolmogorov Smirnov and Shapiro Wilk), Histogram and Normal Q-Q Plot.

Firstly, for faculty member's gender, Skewness and Kurtosis statistics given in Table 4.9. showed that we did not violate normality in male and female groups as all scores are between -3 and +3.

Table 4.9.

Skewness and Kurtosis Statistics for Faculty Member Gender Groups

	Skewness		Kurtosis	
	Statistic	Std. Error	Statistic	Std. Error
Male	-.31	.13	-1.39	.25
Female	-.63	.14	-1.06	.28

The second step in checking normality for gender groups was the tests of normality. As Table 4.10. shows that both tests for both groups have a significance value smaller than .05, and therefore normality assumption is violated.

Table 4.10.

Tests of Normality for Faculty Member's Gender

	Kolmogorov-Smirnov		Shapiro-Wilk	
	Statistic	*p*	Statistic	*p*
Male	.17	.00	.90	.00
Female	.18	.00	.87	.00

Another way to check normality is using histogram curves. Histogram curves for gender presented below show us that not having most scores in the center of distributions, we violated normality assumption in male and female groups.

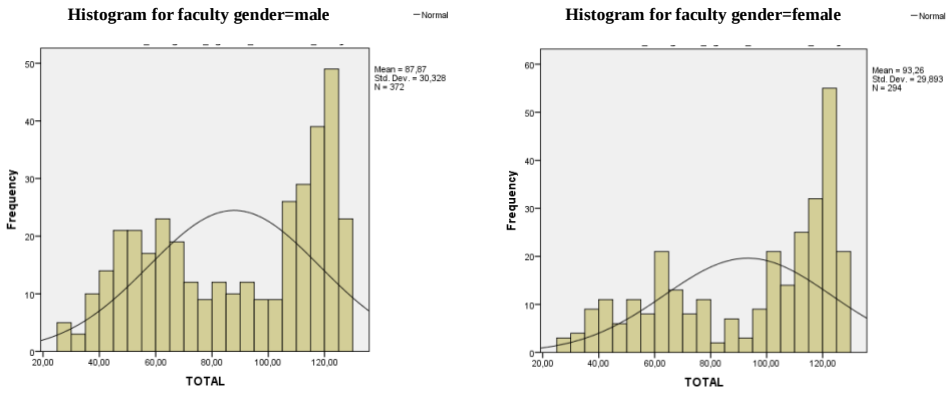

Figure 4.1. Histograms Showing the Distribution of Student Evaluations of Teaching Effectiveness for Male and Female Faculty Members Group Respectively

Lastly, normal Q-Q plot of scores whose figures for male and female groups given below also reveal that not having all dots close to the line, we violated the normality assumption in both groups.

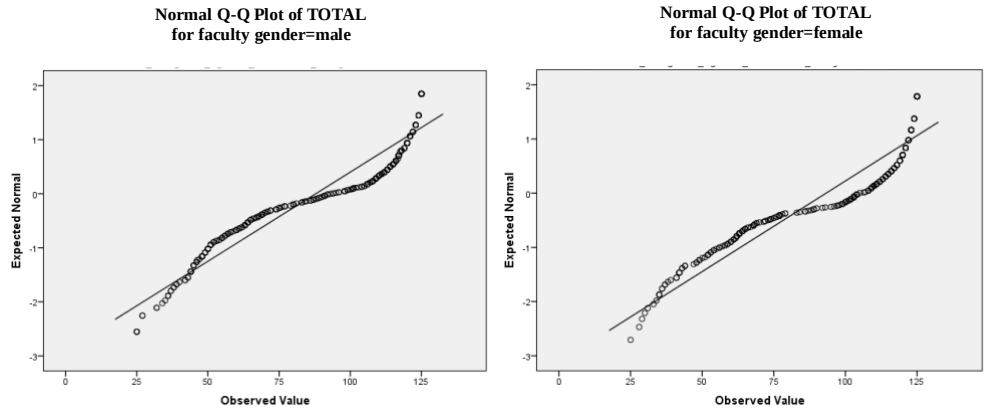

Figure 4.2. Normal Q-Q Plots of Student Evaluations of Teaching Effectiveness for Male and Female Faculty Members Group Respectively

All in all, although normality assumption is violated according to Kolmogorov-Smirnov and Shapiro-Wilk tests, histograms and Q-Q plots, it is acceptable to

continue to the analysis looking at Skewness Kurtosis statistics and our sample size. The central limit theorem tells us no matter how the sample data seems in shape, the estimate will be obtained from a normal distribution in large samples (Field, 2018).

As to normality check for student's gender, Skewness and Kurtosis show that we did not violate the normality assumption for gender groups as all results are between -3 and +3.

Table 4.11.

Skewness and Kurtosis Statistics for Student Gender Groups

	Skewness		Kurtosis	
	Statistic	Std. Error	Statistic	Std. Error
Male student	-.34	.15	-1.24	.29
Female student	-.54	.12	-1.27	.25

As to the tests of normality, both tests for both groups have a significance value lower than .05, which means that normality assumption is violated.

Table 4.12.

Tests of normality for student's gender

	Kolmogorov-Smirnov		Shapiro-Wilk	
	Statistic	*p*	Statistic	*p*
Male	.14	.00	.92	.00
Female	.19	.00	.86	.00

Thirdly, histogram curves for student gender presented below show us that not having most scores in the center of the distributions, we violated normality assumption for both groups.

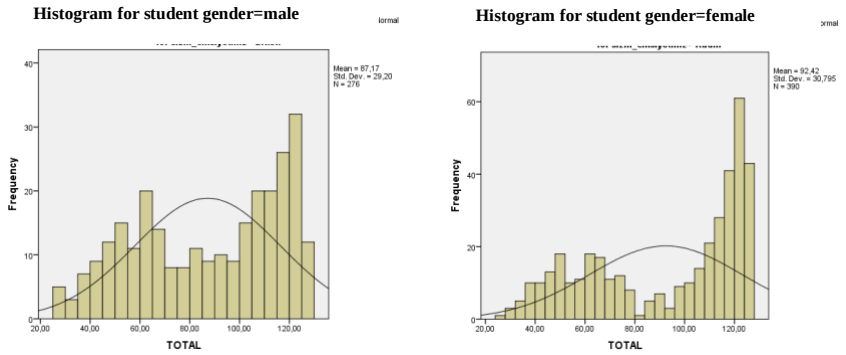

Figure 4.3. Histograms Showing the Distribution of Student Evaluations of Teaching Effectiveness for Male and Female Student Group Respectively

Lastly, normal Q-Q plot of scores whose figures for two different gender groups given below reveal that not having all dots closer to the line, we violated the normality assumption in these groups.

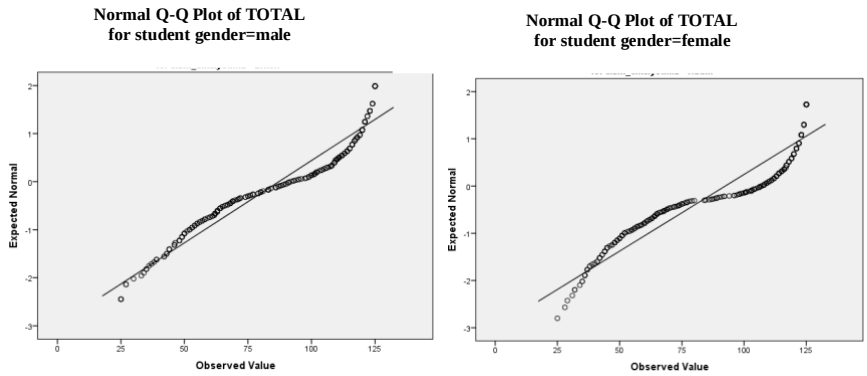

Figure 4.4. Normal Q-Q Plots of Student Evaluations of Teaching Effectiveness for Male and Female Student Group Respectively

Overall, normality assumption is violated according to Kolmogorov-Smirnov and Shapiro-Wilk tests, histograms and Q-Q plots; however, it is acceptable continue to the analysis looking at Skewness Kurtosis statistics and our sample size which can be justified by the central limit theorem as mentioned earlier (Field, 2018).

Lastly, we need to check normality for interaction between faculty member's gender and student's gender. Skewness and Kurtosis statistics given below in Table 4.13. show that we did not violate the normality assumption due to the fact that all values

are between -3 and +3. Kolmogorov-Smirnov and Shapiro-Wilk test statistics are given in Table 4.14. According to these tests, normality is violated in all groups as we have a significance value lower than .05.

Table 4.13.

*Skewness and Kurtosis Statistics for Faculty Member's Gender*Student Gender*

	Skewness		Kurtosis	
	Statistic	Std. Error	Statistic	Std. Error
Male faculty				
Male student	-.27	.18	-1.30	.36
Female student	-.35	.18	-1.47	.35
Female faculty				
Male student	-.47	.25	-1.08	.49
Female student	-.73	.17	-.98	.34

Table 4.14.

*Tests of Normality for Faculty Member's Gender*Student Gender*

	Kolmogorov-Smirnov		Shapiro-Wilk	
	Statistic	*p*	Statistic	*p*
Male faculty				
Male student	.13	.00	.92	.00
Female student	.2	.00	.87	.00
Female faculty				
Male student	.16	.00	.91	.00
Female student	.2	.00	.84	.00

In addition, histogram curves given below for faculty member's gender*student gender presented below show us that we violated normality assumption as we do not have most scores in the center of the distributions.

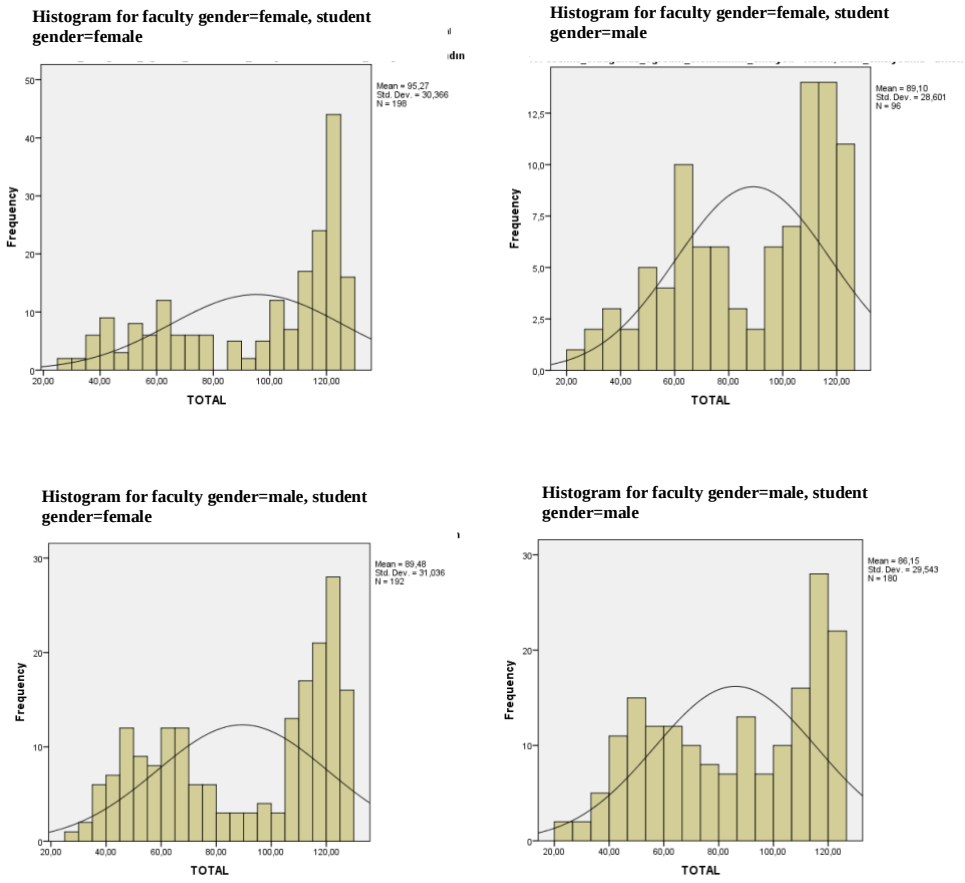

Figure 4.5. Histograms Showing the Distribution of Student Evaluations of Teaching Effectiveness for Female Faculty Female Student Group, Female Faculty Male Student Group, Male Faculty Female Student Group, and Male Faculty Male Student Group Respectively

Furthermore, normal Q-Q plots of student evaluations of teaching effectiveness whose figures for faculty member's gender*student gender given below reveal that not having all dots close to the line we violated the normality assumption in these groups.

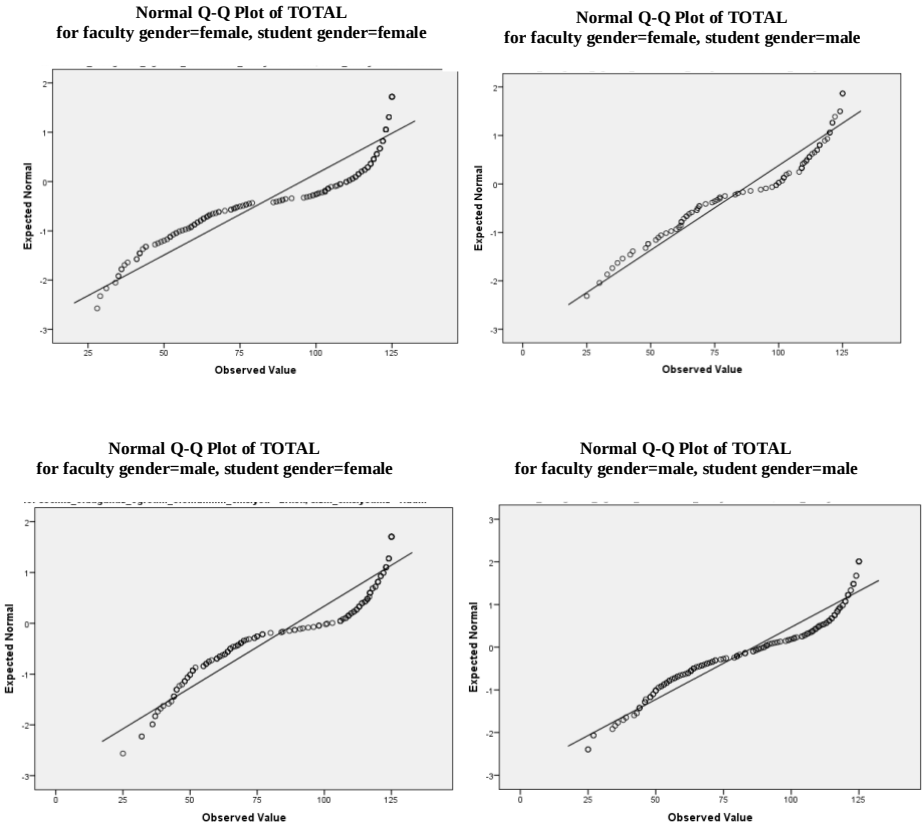

Figure 4.6. Normal Q-Q Plots of Student Evaluations of Teaching Effectiveness for Female Faculty Female Student Group, Female Faculty Male Student Group, Male Faculty Female Student Group and Male Faculty Male Student Group Respectively

All in all, although normality assumption is violated for the interaction between faculty gender and student gender according to Kolmogorov-Smirnov and Shapiro-Wilk tests, histograms and Q-Q plots, it is acceptable to continue to the analysis looking at Skewness Kurtosis statistics and our sample size (Field, 2018).

Another assumption of ANOVA test is homogeneity of variances. Levene's test was used for this purpose and the results showed that we did not violate homogeneity of variances assumption on the grounds that we have a significance value higher than .05, therefore we fail to reject the null hypothesis of the assumption which says that the error variance of the dependent variable is the same in all groups, $F(3,662)=1.66$, $p>.05$.

Table 4.15.

Levene's Test of Equality of Error Variances

F	*df1*	*df2*	*p*
1.66	3	662	.18*

**p*>.05

Since all the assumptions of two-way ANOVA test were checked, we could continue with the results of the test.

4.2.1.2 Interpretation of two-way ANOVA

A 2*2 ANOVA was conducted on student evaluations of teaching effectiveness to determine if the results differ by faculty member's gender and/or student's gender.

The null hypothesis for the main effect of faculty member's gender factor which involves the comparison of two levels of male and female states that there is no difference between these two levels, in other words, faculty member's gender has no effect on student evaluations of teaching effectiveness, $\mu_{malefaculty}=\mu_{femalefaculty}$ in statistical terms. The alternative hypothesis for the main effect of gender is that there is a difference in student evaluations of teaching effectiveness between two levels of faculty member's gender, $\mu_{malefaculty} \neq \mu_{femalefaculty}$ in statistical terms. The null hypothesis for the main effect of student's gender factor which has also two levels is that there is no difference between these two levels, in other words, student's gender has no effect on student evaluations of teaching effectiveness, $\mu_{studentmale}=\mu_{studentfemale}$ in statistical terms. The alternative hypothesis for the main effect of student's gender factor is that there is a difference between two levels of student's gender, $\mu_{studentmale}\neq\mu_{studentfemale}$ in statistical terms. As to the null hypothesis for the interaction between faculty member's gender and student's gender, the null hypothesis states that there is no interaction between factors faculty member's gender and student's gender, which means the effect of faculty member's gender does not depend on the effect of student's gender and vice versa. The alternative interaction hypothesis states that there is an interaction between faculty member's gender and student gender and

main effects cannot predict the mean differences between groups. The study has two factors (independent variables). The first independent variable, faculty member's gender, has two levels: Male and female. The second independent variable, student's gender, has also two levels: Male and female. The dependent variable is student evaluations of teaching effectiveness.

The results did not show a statistically significant interaction between faculty member's gender and student gender, $F(1,662)=.34$, $p>.05$. As there is no interaction between faculty member's gender and student gender, we look at the main effects. However, no significant main effect was found for faculty member's gender, $F(1,662)=3.21$, $p>.05$ and for student gender $F(1,662)=3.81$, $p>.05$.

Table 4.16.

*2*2 ANOVA Summary Table*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
Faculty gender	2909.32	1	2909.32	3.21	.07	.01
Student gender	3447.04	1	3447.04	3.81	.05	.01
Faculty gender*Student gender	306.82	1	306.82	.34	.56	.00
Error	599566.85	662	905.69			
Corrected Total	607826.20	665				

As a result, we fail to reject the null hypothesis for research question 1. This analysis shows us that there is no statistically significant difference in the scores of student evaluations of male and female faculty members' teaching effectiveness, and student evaluations of teaching does not also significantly differ either by students' gender or by the interaction between faculty member's gender and student gender.

4.2.2 Inferential statistics for research question 2

This study was conducted to find out whether student evaluations of teaching effectiveness scores differ by students' grades taken at the end of the semester. Therefore, the second research question of this study was "Is there a statistically significant difference in student evaluations of teaching effectiveness by student's grade?".

4.2.2.1 Assumptions of one-way ANOVA

The independent measures ANOVA test has three assumptions which are independent observations, normality and homogeneity of variances; and therefore, before running the test these assumptions were checked.

Firstly, the independent observations assumption is violated because as mentioned earlier the data were collected in participants' classrooms. Any interaction between students to choose the same faculty member to evaluate may have caused this violation.

As to the normality check, four tests were conducted: Skewness-Kurtosis, Kolmogorov-Smirnov and Shapiro-Wilk, histogram and normal Q-Q plot of scores. Table 4.17. shows the results for Skewness and Kurtosis for all groups. To have a normal distribution, we expect the results to be close to zero, between the values of -3 and +3. As all the results listed in this table are between these values, normality assumption is not violated.

Table 4.17.

Skewness and Kurtosis Statistics for Grade Groups

	Skewness		Kurtosis	
	Statistic	Std. Error	Statistic	Std. Error
AA	-1.77	.17	2.1	.35
BA	-1.38	.24	.83	.48
BB	-.53	.25	-1.04	.49
CB	.08	.29	-1.24	.57
CC	.73	.30	-1.30	.59
DC	.73	.39	-.80	.76
DD	.58	.39	-.80	.76
FD	1.28	.43	2.78	.83
FF	.85	.52	.93	1.01
S	-1.08	1.01	1.75	2.62
W	-.48	1.01	-2.35	2.62
NA	1.28	.79	1.81	1.59

The second step in normality check is Kolmogorov-Smirnov and Shapiro Wilk tests given in Table 4.18. For the grade groups AA, BA, BB, CB, DC, DD; as *p* value is below .05, normality assumption is violated. However, for CC and FD groups, normality is not violated according to Kolmogorov-Smirnov test, but it is violated according to Shapiro-Wilk test. For FF and NA groups, normality is not violated according to both tests.

Table 4.18.

Tests of Normality for Grade Groups

	Kolmogorov-Smirnov		Shapiro-Wilk	
	Statistic	*p*	Statistic	*p*
AA	.23	.00	.72	.00
BA	.2	.00	.79	.00
BB	.15	.00	.91	.00
CB	.13	.01	.93	.00
CC	.1	.18	.94	.00
DC	.19	.00	.88	.00
DD	.17	.01	.92	.01
FD	.14	.15	.91	.01
FF	.15	.2	.95	.39
S	.26	.	.94	.64
W	.24	.	.94	.67
NA	.21	.2	.89	.28

In addition, histogram curves given below for grade groups show us that we violated normality assumption as we do not have most scores in the center of the distributions.

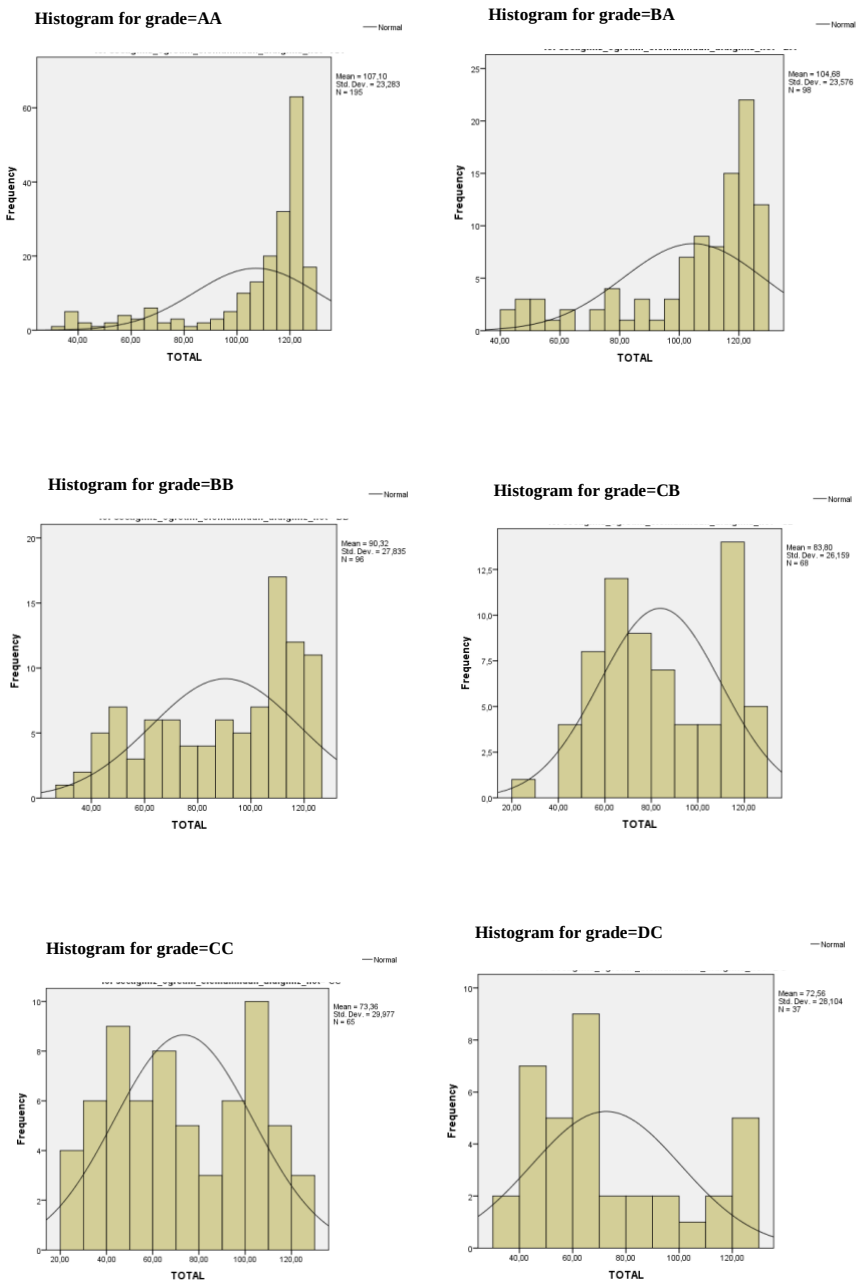

Figure 4.7. Histograms Showing the Distribution of Student Evaluations of Teaching Effectiveness for Grade AA, BA, BB, CB, CC, DC, DD, FD, FF, S, W, NA Groups

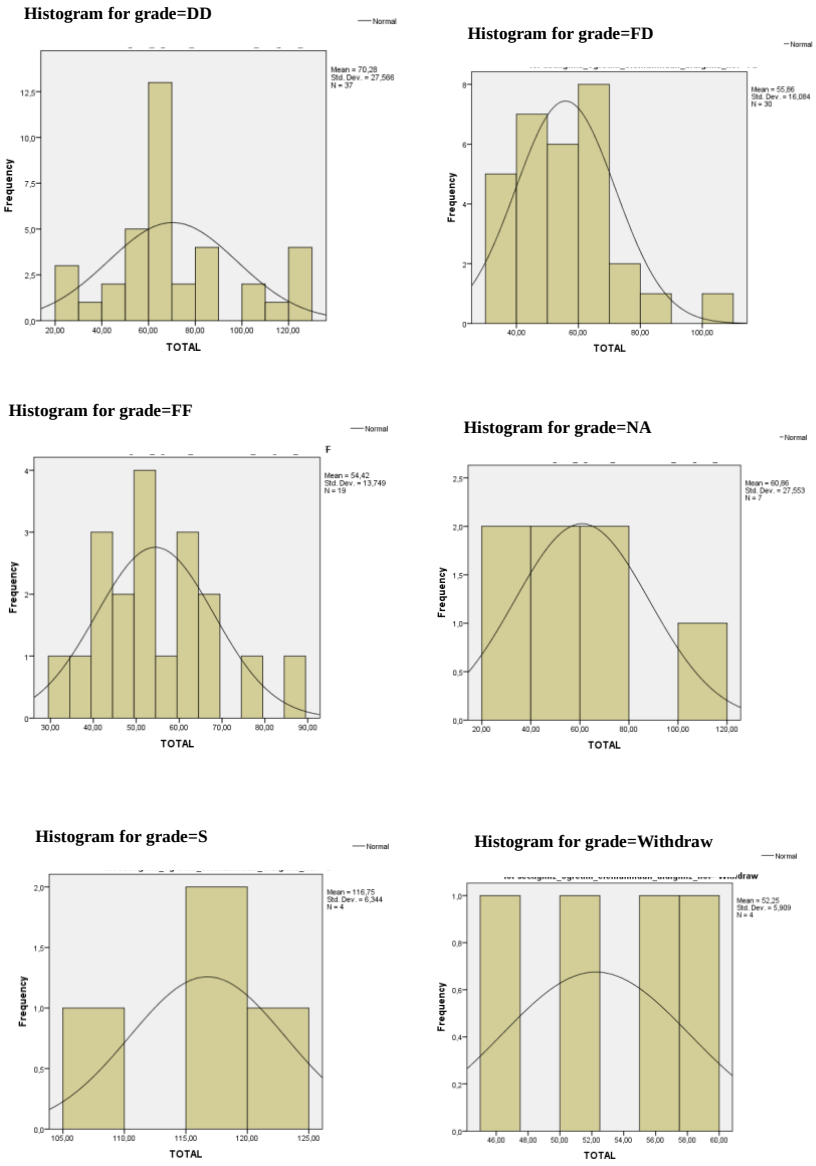

Figure 4.7. (cont'd) Histograms Showing the Distribution of Student Evaluations of Teaching Effectiveness for Grade AA, BA, BB, CB, CC, DC, DD, FD, FF, S, W, NA Groups

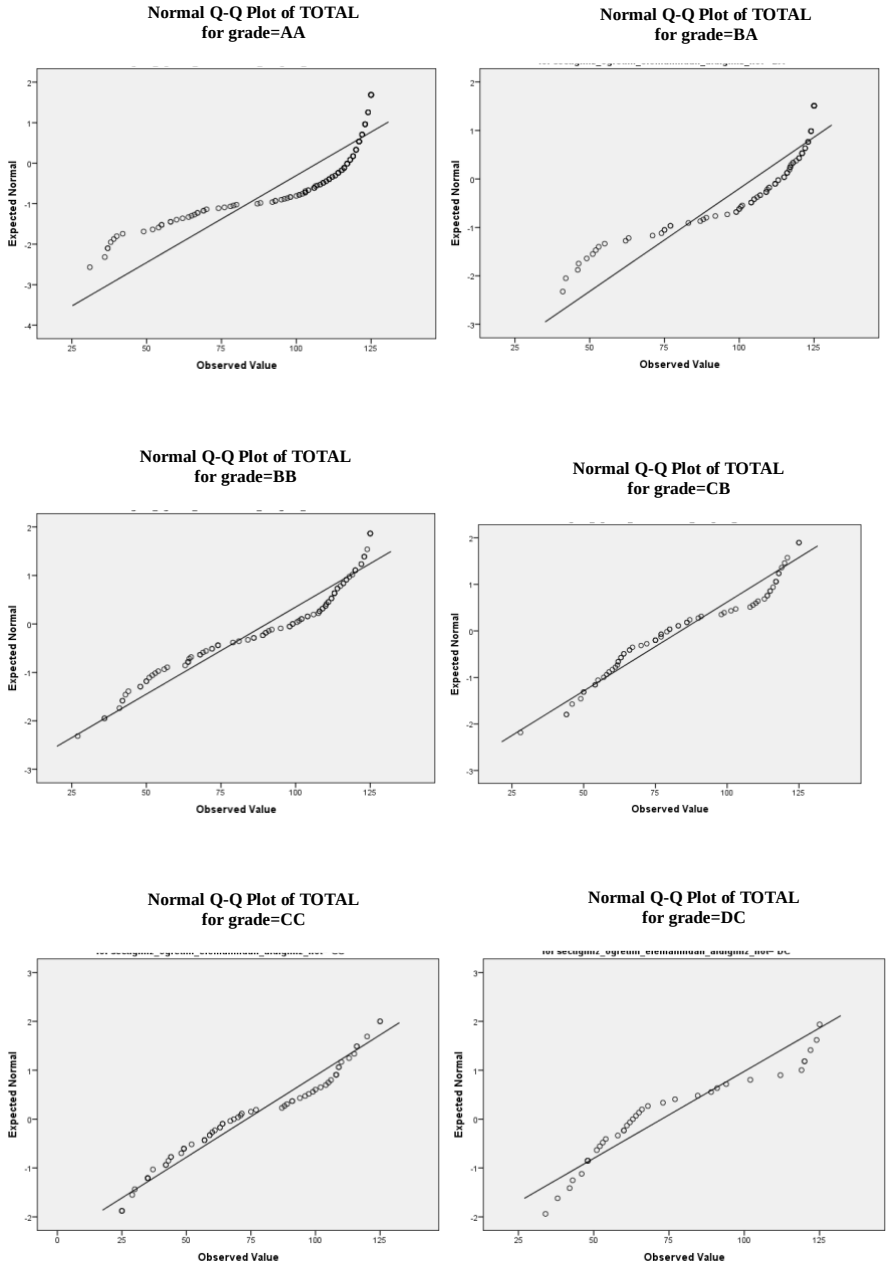

Figure 4.8. Normal Q-Q Plots of Student Evaluations of Teaching Effectiveness for Grade AA, BA, BB, CB, CC, DC, DD, FD, FF, S, W, NA Groups

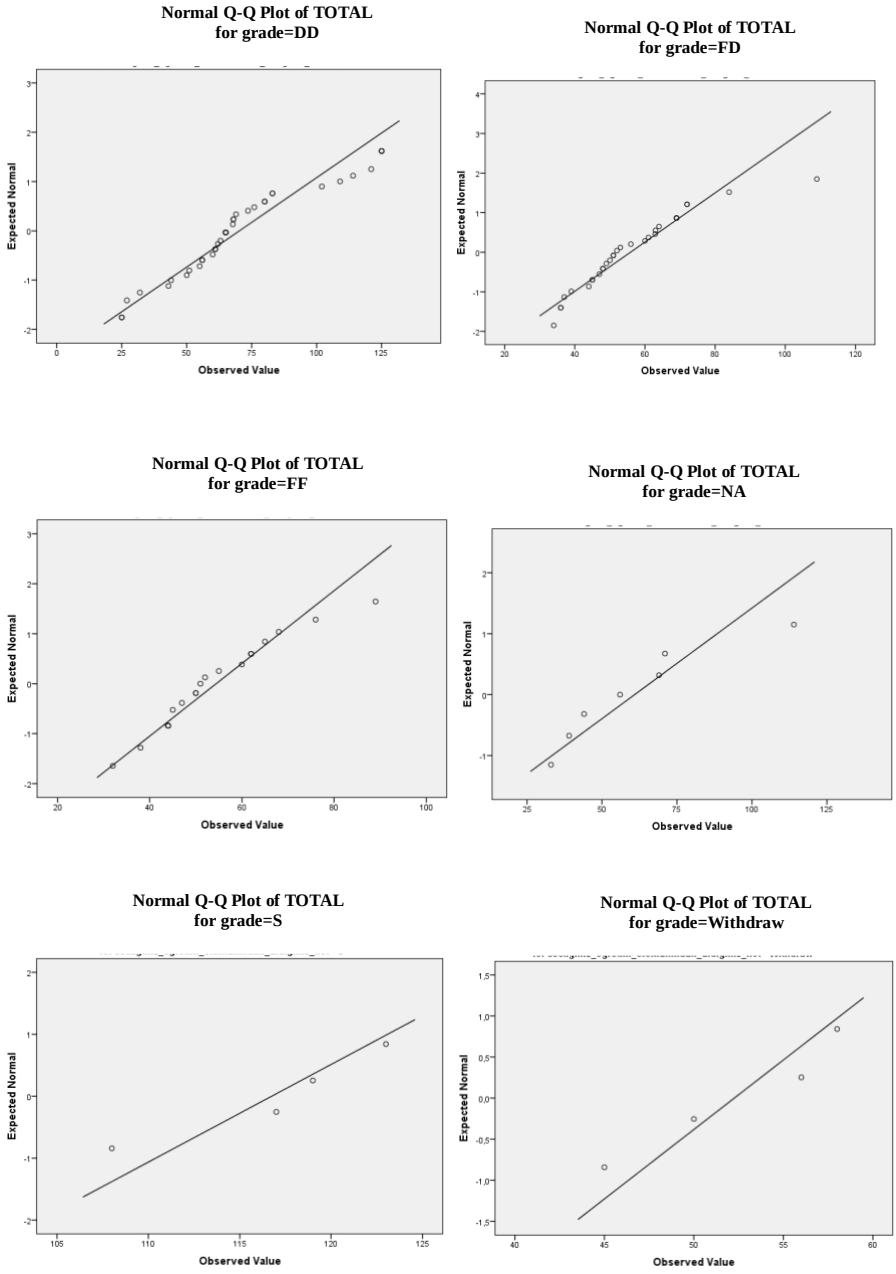

Figure 4.8. (cont'd) Normal Q-Q Plots of Student Evaluations of Teaching Effectiveness for Grade AA, BA, BB, CB, CC, DC, DD, FD, FF, S, W, NA Groups

According to Normal Q-Q Plots, it is clear that normality assumption is violated. Although normality is violated according to normality tests, histogram and normal Q-Q plots, it is acceptable to continue our analysis based on Skewness Kurtosis statistics and our sample size.

Another assumption of one-way ANOVA test is homogeneity of variances. Levene's test was used for this purpose and the results of the test showed that homogeneity of variances assumption is violated on the grounds that we have a significance value higher than .05; therefore we reject the null hypothesis of the assumption which says that error variance of dependent variable is the same in all groups, $F(11, 648)=5.50$, $p<.05$. As this assumption is violated, Welch's F test is used for the analysis instead of one-way ANOVA.

Table 4.19.

Levene's Test of Equality of Error Variances

F	*df1*	*df2*	*p*
5.50	11	648	.00*

*$p<.05$

4.2.2.2 Interpretation of Welch's F test

For the second research question, the independent variable which is grade has 12 levels: AA, BA, BB, CB, CC, DC, DD, FD, FF, Withdraw (W), Successful (S) and NA. The dependent variable is student evaluations of teaching effectiveness.

The null hypothesis states that all 12 groups of grade have the same population mean, $\mu_{AA}=\mu_{BA}=\mu_{BB}=\mu_{CB}=\mu_{CC}=\mu_{CD}=\mu_{DD}=\mu_{FD}=\mu_{FF}=\mu_{W}=\mu_{S}=\mu_{NA}$ in statistical terms. The alternative hypothesis states that at least one mean difference among populations exists.

The findings of the study showed that student evaluations of teaching effectiveness significantly differ by student's grade *Welch's* $F(11, 52.65) = 55.79$, $p=.00$. Scheffe post hoc test showed that the grade group AA ($M=107.10$, $SD=23.28$) significantly differed from grade group BB ($M=90.32$, $SD=27.83$), CB ($M=83.80$, $SD=26.16$), CC ($M=73.36$, $SD=29.98$), DC ($M=72.56$, $SD=28.10$), DD ($M=70.28$, $SD=27.57$), FD ($M=55.86$, $SD=16.08$), FF ($M=54.42$, $SD=13.75$) and NA ($M=60.86$, $SD=27.55$). Grade group BA ($M=104.68$, $SD=23.58$) significantly differed from grade groups CB

(*M*=83.80, *SD*=26.16), CC (*M*=73.36, *SD*=29.98), DC (*M*=72.56, *SD*=28.10), DD (*M*=70.28, *SD*=27.57), FD (*M*=55.86, *SD*=16.08), FF (*M*=54.42, *SD*=13.75), and NA (*M*=60.86, *SD*=27.55). BB (*M*=90.32, *SD*=27.83) group statistically differed from FD (*M*=55.86, *SD*=16.08), FF (*M*=54.42, *SD*=13.75). CB (*M*=83.80, *SD*=26.16) statistically differed from FD (*M*=55.86, *SD*=16.08) and FF (*M*=54.42, *SD*=13.75). FD (*M*=55.86, *SD*=16.08) statistically differed from S (*M*=116.75, *SD*=6.34). S grade group (*M*=116.75, *SD*=6.34) significantly differed from FF (*M*=54.42, *SD*=13.75).

Overall, the results of the Welch's F test and Scheffe post hoc test revealed a statistically significant main effect of grade on student evaluations of teaching effectiveness scores. Therefore, the null hypothesis is rejected for this research question.

Table 4.20.

Scheffe post hoc test

Grade	Grade	MD	SD	p
	BA	2.42	3.11	1
	BB	16.78*	3.13	0
	CB	23.30*	3.53	0
	CC	33.74*	3.59	0
	DC	34.55*	4.5	0
AA	DD	36.82*	4.5	0
	FD	51.24*	4.92	0
	FF	52.68*	6.03	0
	NA	46.25*	9.65	0.02
	S	-9.65	12.67	1
	W	54.85	12.67	0.07

Table 4.20. (cont'd)

Grade	Grade	MD	SD	p
	AA	-2.42	3.11	1
	BB	14.36	3.6	0.15
	CB	20.87*	3.96	0
	CC	31.32*	4.01	0
	DC	32.12*	4.84	0
BA	DD	34.40*	4.84	0
	FD	48.81*	5.23	0
	FF	50.26*	6.29	0
	NA	43.82*	9.81	0.048
	S	-12.07	12.79	1
	W	52.43	12.79	0.12
	AA	-16.78*	3.13	0
	BA	-14.36	3.6	0.15
	CB	6.52	3.98	0.1
	CC	16.96	4.03	0.1
	DC	17.77	4.85	0.27
BB	DD	20.04	4.85	0.11
	FD	34.46*	5.25	0
	FF	35.90*	6.3	0
	NA	29.47	9.82	0.62
	S	-26.43	12.8	0.96
	W	38.07	12.8	0.64
	AA	-23.30*	3.53	0
	BA	-20.87*	3.96	0
	BB	-6.52	3.98	0.99
	CC	10.45	4.35	0.89
	DC	11.25	5.12	0.94
CB	DD	13.53	5.12	0.8
	FD	27.94*	5.5	0.01
	FF	29.38*	6.51	0.04
	NA	22.95	9.95	0.91
	S	-32.95	12.9	0.84
	W	31.55	12.9	0.87

Table 4.20. (cont'd)

Grade	Grade	MD	SD	p
CC	AA	-33.74*	3.59	0
	BA	-31.32*	4.01	0
	BB	-16.96	4.03	0.09
	CB	-10.45	4.35	0.89
	DC	0.8	5.17	1
	DD	3.08	5.17	1
	FD	17.49	5.54	0.53
	FF	18.94	6.54	0.68
	NA	12.5	9.98	1
	S	-43.39	12.92	0.42
	W	21.11	12.92	0.99
DC	AA	-34.55*	4.5	0
	BA	-32.12*	4.84	0
	BB	-17.77	4.85	0.27
	CB	-11.25	5.12	0.94
	CC	-0.8	5.17	1
	DD	2.28	5.83	1
	FD	16.69	6.16	0.77
	FF	18.13	7.08	0.83
	NA	11.7	10.34	1
	S	-44.2	13.2	0.43
	W	20.31	13.2	0.1
DD	AA	-36.82*	4.5	0
	BA	-34.40*	4.84	0
	BB	-20.04	4.85	0.11
	CB	-13.53	5.12	0.8
	CC	-3.08	5.17	1
	DC	-2.28	5.83	1
	FD	14.42	6.16	0.91
	FF	15.86	7.08	0.93
	NA	9.42	10.34	1
	S	-46.47	13.2	0.34
	W	18.03	13.2	0.1

Table 4.20. (cont'd)

Grade	Grade	MD	SD	p
FD	AA	-51.24*	4.92	0
	BA	-48.81*	5.23	0
	BB	-34.46*	5.25	0
	CB	-27.94*	5.5	0.01
	CC	-17.49	5.54	0.53
	DC	-16.69	6.16	0.77
	DD	-14.42	6.16	0.91
	FF	1.44	7.35	1
	NA	-4.99	10.53	1
	S	-60.89*	13.35	0.04
	W	3.61	13.35	1
FF	AA	-52.68*	6.03	0
	BA	-50.26*	6.29	0
	BB	-35.90*	6.3	0
	CB	-29.38*	6.51	0.04
	CC	-18.94	6.54	0.68
	DC	-18.13	7.08	0.83
	DD	-15.86	7.08	0.93
	FD	-1.44	7.35	1
	NA	-6.44	11.09	1
	S	-62.33*	13.8	0.04
	W	2.17	13.8	1
NA	AA	-46.25*	9.65	0.02
	BA	-43.82*	9.81	0.048
	BB	-29.47	9.82	0.62
	CB	-22.95	9.95	0.91
	CC	-12.5	9.98	1
	DC	-11.7	10.34	1
	DD	-9.42	10.34	1
	FD	4.99	10.53	1
	FF	6.44	11.09	1
	S	-55.9	15.72	0.32
	W	8.61	15.72	1

Table 4.20. (cont'd)

Grade	Grade	MD	SD	p
	AA	9.65	12.67	1
	BA	12.07	12.79	1
	BB	26.43	12.8	0.96
	CB	32.95	12.9	0.84
	CC	43.39	12.92	0.42
S	DC	44.2	13.2	0.43
	DD	46.47	13.2	0.34
	FD	60.89*	13.35	0.04
	FF	62.33*	13.8	0.04
	NA	55.89	15.72	0.32
	W	64.5	17.73	0.28
	AA	-54.85	12.67	0.07
	BA	-52.43	12.79	0.12
	BB	-38.07	12.8	0.64
	CB	-31.55	12.9	0.87
	CC	-21.11	12.92	0.99
W	DC	-20.31	13.2	0.1
	DD	-18.03	13.2	0.1
	FD	-3.61	13.35	1
	FF	-2.17	13.8	1
	NA	-8.61	15.72	1
	S	-64.5	17.73	0.28

4.2.3 Inferential statistics for research question 3

The third research question of the study is "Is there a statistically significant difference in student evaluations of teaching effectiveness by student's discipline?" In this study, the independent variable is student's discipline and the dependent variable is student evaluations of teaching effectiveness. The independent variable has 6 levels: Architecture, Humanities, Engineering, Natural Sciences, Economics and Administrative Sciences, and Educational Sciences; therefore, one-way ANOVA test will be conducted.

4.2.3.1 Assumptions of one-way ANOVA

Before conducting one-way ANOVA test to understand if there is a statistically significant difference in student evaluations of teaching effectiveness by students' discipline, it is necessary to check the assumptions of the test which are independent observations, normality and homogeneity of variances.

The independent observations assumption is violated because the instrument was implemented in participants' classrooms, so each individual may have been affected by other participants. As to the normality check, four tests were conducted: Skewness and Kurtosis, Kolmogorov-Smirnov and Shapiro-Wilk, histogram and normal Q-Q plot of scores.

Table 4.21.

Skewness and Kurtosis Statistics for 6 Disciplines

	Skewness		Kurtosis	
	Statistic	Std. Error	Statistic	Std. Error
Architecture	.13	.28	-1.57	.55
Humanities	-1.05	.30	-.22	.60
Engineering	-.53	.17	-1.12	.33
Natural Sci.	-.05	.24	-1.53	.47
Eco. & Adm. Sci.	-.88	.25	-.48	.50
Educational Sci.	-.48	.23	-1.39	.45

Since all Skewness-Kurtosis statistics given in the table above are between -3 and +3, it is clear that the normality assumption is not violated.

Table 4.22.

Tests of Normality for Disciplines

	Kolmogorov-Smirnov		Shapiro-Wilk	
	Statistic	*p*	Statistic	*p*
Architecture	.15	.00	.89	.00
Humanities	.24	.00	.81	.00
Engineering	.16	.00	.89	.00
Natural Sci.	.15	.00	.90	.00
Eco. & Adm. Sci.	.2	.00	.86	.00
Educational Sci.	.21	.00	.85	.00

As to Kolmogorov-Smirnov and Shapiro Wilk tests given in Table 4.22.; as mentioned earlier, when all significance values are below .05, it means that normality assumption is violated. Also, histograms in Figure 4.9 show that not having most scores in the center of the distribution, we violated normality assumption.

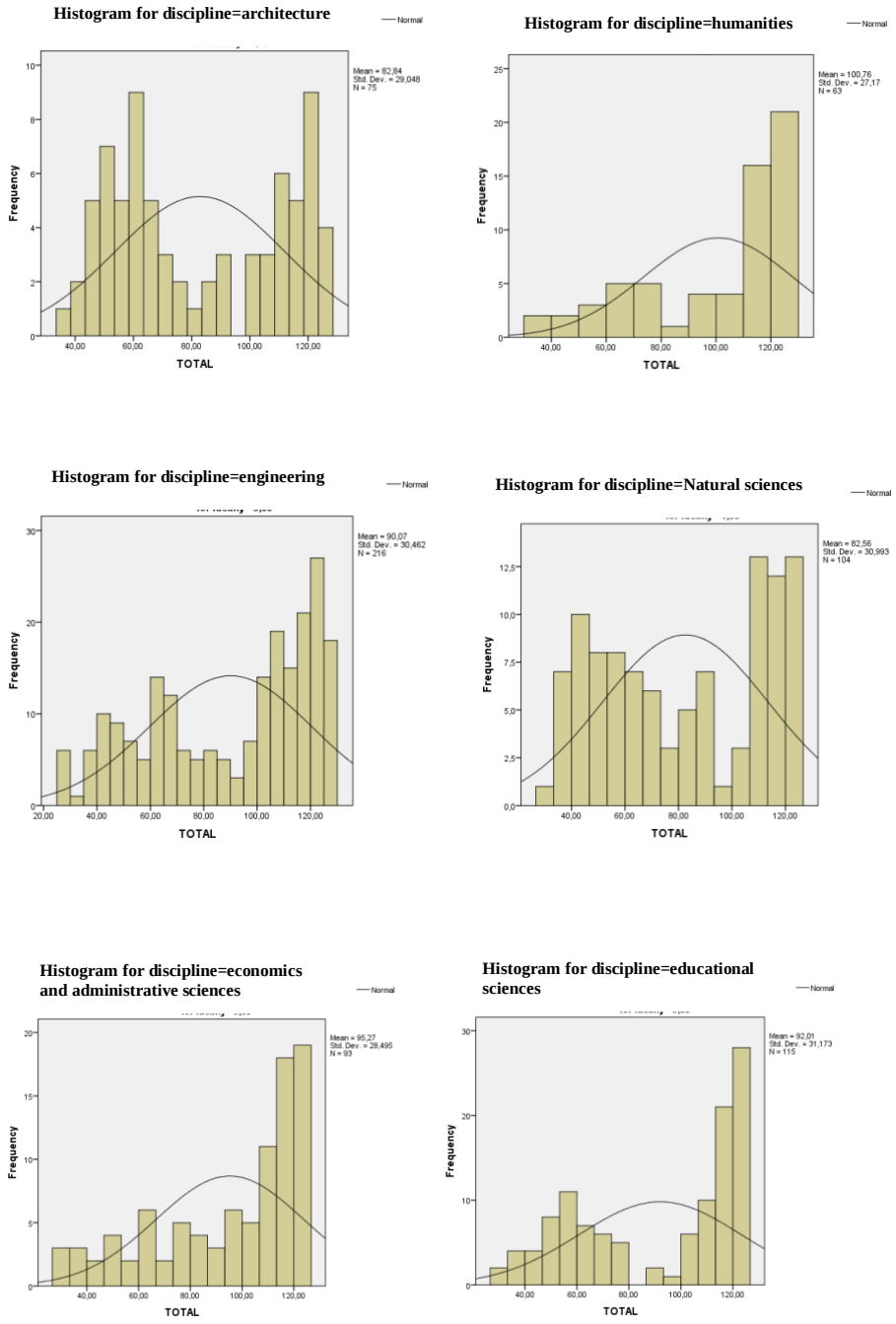

Figure 4.9. Histogram Showing the Distribution of Student Evaluations of Teaching Effectiveness for Architecture, Humanities, Engineering, Natural Sciences, Economics and Administrative Sciences, and Educational Sciences Group

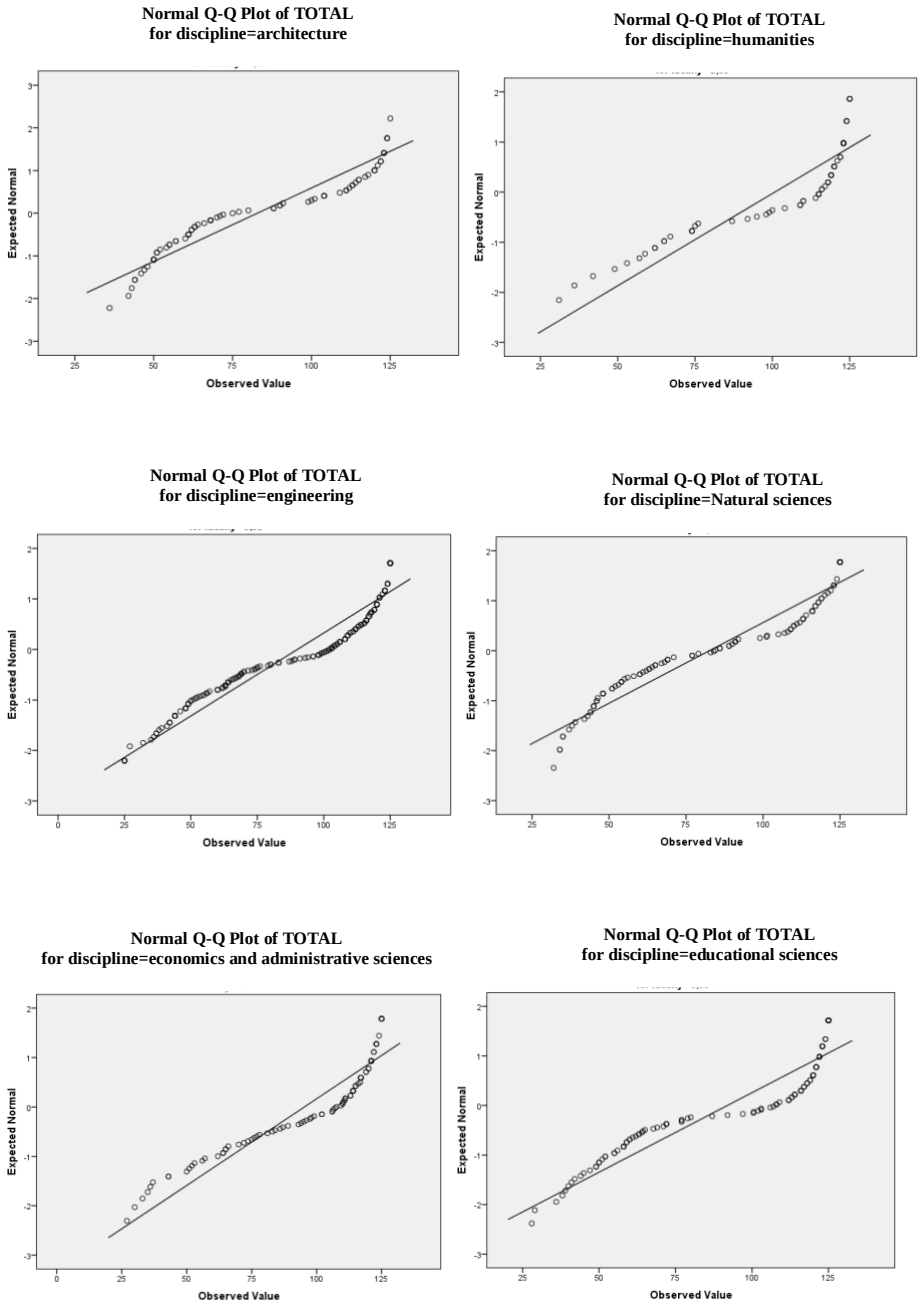

Figure 4.10. Normal Q-Q Plot of Student Evaluations of Teaching Effectiveness for Architecture, Humanities, Engineering, Natural Sciences, Economics and Administrative Sciences, and Educational Sciences Group

Lastly, as shown in Figure 4.10, as not all dots are close to the line in these plots, normality assumption is violated. All in all, although the normality assumption is violated according to normality tests and histograms, normal Q-Q plots, it is

acceptable to continue with one-way ANOVA because of Skewness Kurtosis statistics and our sample size.

Another assumption of one-way ANOVA test is homogeneity of variances. Levene's test was used for this purpose and the results of the test showed that homogeneity of variances assumption is violated on the grounds that we have a significance value higher than .05; therefore we reject the null hypothesis of the assumption which says that error variance of dependent variable is the same in all groups, $F(5, 660)=2.52$, $p<.05$. As this assumption is violated, Welch's F test is used for the analysis instead of one-way ANOVA.

Table 4.23.

Levene's Test of Equality of Error Variances

F	*df1*	*df2*	*p*
2.52	5	660	.03*

*$p<.05$

4.2.3.2 Interpretation of Welch's F test

The null hypothesis states that all six groups of discipline have the same population mean, $\mu_{architecture}=\mu_{humanities}=\mu_{engineering}=\mu_{naturalsci}=\mu_{administrativesci}=\mu_{education}$ in statistical terms. The alternative hypothesis states that at least one mean difference among the populations exists.

The student evaluations of teaching effectiveness significantly differ by student's discipline *Welch's* $F(5, 251.04) = 4.74$, $p=.00$. Scheffe post hoc test showed that the Architecture group ($M=82.84$, $SD=29.05$) significantly differed from grade group Humanities ($M=100.76$, $SD=27.17$), and Humanities group ($M=100.76$, $SD=27.17$) significantly differed from grade group Natural Sciences ($M=82.56$, $SD=30.99$).

Table 4.24.

Scheffe post hoc test

Discipline	Discipline	MD	SD	p
Architecture	Humanities	-17.92*	5.12	0.03
	Engineering	-7.23	4.01	0.66
	Natural Sciences	0.28	4.54	1.00
	Eco. & Adm. Sciences	-12.43	4.65	0.21
	Educational Sciences	-9.17	4.45	0.51
Humanities	Architecture	17.92*	5.12	0.03
	Engineering	10.69	4.29	0.29
	Natural Sciences	18.20*	4.78	0.01
	Eco. & Adm. Sciences	5.49	4.89	0.94
	Educational Sciences	8.75	4.69	0.63
Engineering	Architecture	7.23	4.01	0.66
	Humanities	-10.69	4.29	0.29
	Natural Sciences	7.51	3.57	0.49
	Eco. & Adm. Sciences	-5.20	3.71	0.86
	Educational Sciences	-1.94	3.46	1.00
Natural Sciences	Architecture	-0.28	4.54	1.00
	Humanities	-18.20*	4.78	0.01
	Engineering	-7.51	3.57	0.49
	Eco. & Adm. Sciences	-12.71	4.27	0.12
	Educational Sciences	-9.45	4.05	0.37
Eco. & Adm. Sciences	Architecture	12.43	4.65	0.21
	Humanities	-5.49	4.89	0.94
	Engineering	5.20	3.71	0.86
	Natural Sciences	12.71	4.27	0.12
	Educational Sciences	3.26	4.18	0.99
Educational Sciences	Architecture	9.17	4.45	0.51
	Humanities	-8.75	4.69	0.63
	Engineering	1.94	3.46	1.00
	Natural Sciences	9.45	4.05	0.37
	Eco. & Adm. Sciences	-3.26	4.18	0.99

4.3 Summary of the Findings

The first research question of the study was “Is there a statistically significant difference between male and female faculty members’ teaching effectiveness scores by students’ and faculty members’ gender?”. This question had two factors (independent variables). The first one was faculty gender, and it had two levels: Male and female. The second independent variable, student gender, also had two levels: Male and female. The dependent variable was student evaluations of teaching effectiveness. The results of two-way ANOVA for the first question did not indicate a statistically significant interaction between faculty members’ gender and students’ gender. Since there was no interaction between faculty gender and student gender, main effects of each were investigated. However, the results did not show any significant effect for faculty gender or for student gender.

The second research question in this study was “Is there a statistically significant difference in student evaluations of faculty teaching effectiveness by student’s grade?”. The independent variable for this question was grade and it had 12 levels: AA, BA, BB, CB, CC, DC, DD, FD, FF, Withdraw (W), Successful (S), and NA. Although one-way ANOVA test was planned, Welch’s F test was used as homogeneity of variances assumption was violated. The results indicated that student evaluations of teaching effectiveness significantly differ by student’s grade.

The third research question was “Is there a statistically significant difference in student evaluations of faculty teaching effectiveness by student’s discipline?”. This research question had one independent variable, discipline, which had 6 levels: Architecture, Humanities, Engineering, Natural Sciences, Economics and Administrative Sciences, and Educational Sciences. For this question although one-way ANOVA test was planned, Welch’s F test was used as homogeneity of variances assumption was violated. The results indicated that the student evaluations of teaching effectiveness scores significantly differ by student’s discipline.

Going back to the main research question of this study, which was "Are there statistically significant differences in scores of student evaluations of teaching effectiveness by faculty members' gender and student characteristics?", it is clear from the analyses mentioned above that the scores of student evaluations of teaching effectiveness did not significantly differ by faculty member's gender and student gender. However, as to the student characteristics, grade and discipline, the scores were found to significantly differ by them.

CHAPTER 5

DISCUSSION AND IMPLICATIONS

This chapter presents a discussion of the results, implications and recommendations for future research.

5.1 Discussion

In this part, a discussion of the results of the study and the possible explanations for these findings are presented. In the first section, it provides a discussion of why student evaluations of teaching effectiveness do not differ when faculty member's gender, student gender and the interaction between them are taken into consideration. In addition, in the second and third sections, the discussions aim to provide explanations for variables grade and discipline in which the study has revealed significant differences.

As the first significant result of the present study, it did not reveal any significant difference as to the student evaluations of teaching effectiveness scores when faculty member's gender, student gender and the interaction between them are considered. In the literature of SETs, while there are many studies which have found the opposite results, there are also studies which deny gender bias in the implementation and interpretation of SETs.

The findings of this study are in line with Fernángez and Mateo (1997) who did not find any statistically significant interaction between faculty member's gender and student gender after analyzing the data from 1,304 students in Madrid, Spain. Similar results were also found in Wheeless and Potorti (1989). In this study, masculinity and

femininity as personality characteristics was found to be more important than whether the faculty member and the student was female or male. Therefore, we can conclude that masculinity and femininity as personality characteristics may be more important for students; and according to students, a good combination of both traits may define effective teaching.

Apart from this, there may be other explanations to the present findings. Firstly, the participants of this study are studying in different departments at Middle East Technical University, which has an international campus culture, has English as the medium of instruction, has many opportunities for international cultural exchanges. This international environment which is welcoming for diversity may mean that students' perceptions of gender stereotypes are not as stereotypical as the perceptions of the research population. Therefore, this multicultural environment may have had impacts on the findings of the study.

Secondly, the numerical visibility of women in teaching positions in Turkey may be another reason for this finding. In the literature, this is discussed as "feminization of teaching". Erginer and Saklan (2016) define feminization of teaching in Turkey as a rise in the number of women as teachers and educators. According to Erginer and Saklan's findings, teaching profession in Turkey tends to become a female job in that the number of female preservice teachers has increased in faculties of education between the years 1997 and 2006. The main reason for this increase, according to the authors, is the compatibility of teaching and domestic responsibilities culturally assigned to women. Apart from this, there are other explanations in the literature about the numerical visibility of women in higher education. Firstly, Danyal-Köker (1988) claims that in the West it has taken long years and efforts to deconstruct the belief that science can only be done by men. However, according to Danyal-Köker, Turkish universities have not gone through the elitist-traditional institutionalization processes unlike their Western colleagues, and this has become an advantage for women who wanted to pursue academic careers. Another advantage of women in Turkey who have followed academic careers was the lack of qualified workforce to

work at universities. Danyal-Köker (1988) argues that the young Turkish Republic needed people who were loyal to revolutions and qualified enough to contribute to the scientific development of the country. However, numerical inadequacy of this workforce allowed women to be employed as academic personnel. Similarly, another argument to explain women's numerical presence in Turkish higher education is put forward by Akbulut (2011) suggesting that as men leave academic positions for better employment and financial gain opportunities outside the university, women take over these academic positions and continue to stay in academia. In addition, Özkanlı (2007) claimed that The Council of Higher Education in Turkey, namely YÖK, has transparent employment procedures and regulations, and this has also contributed to women's advancement at universities. According to Özbilgin and Healy (2004), this standardization in recruitment and appointment offers women higher chances when access to full professorship is considered while this is not the case in many parts of Western Europe and North America where it is not common to find standardization in recruitment and appointment.

To conclude what we have discussed about the first finding of the study, although universities are gendered institutions as put forward by Acker (1990), with the absence of women in leadership and decision-making positions and with expected gendered behaviors from men and women; the study did not find statistically significant differences in student evaluations of teaching effectiveness by faculty member's gender and student's gender. Despite underlying inequalities, the numerical visibility of women in teaching position in higher education and leading events to this visibility may be the reasons for this finding in this study. In addition, students may be assigning femininity and masculinity as personality characteristics rather than gender traits as shown by Wheeless and Potorti (1989). In addition, the multicultural campus environment of Middle East Technical University which appreciates diversity may be reducing the impacts of student's gender stereotypes.

The second important finding of this research is that it revealed statistically significant differences between groups of students of by their grades. Grade here, as

mentioned earlier, refers to students' final course grade in the selected faculty member's course. This result is in line with Lin (2015). In Lin's study, there was a statistically significant correlation between students' grade and SET scores after students received grade feedback before the implementation of SET. However, when grade knowledge was not available to students before the administration of the instrument, there was not an association between students' course grade and SET scores. As mentioned this is in line with the findings of the present research in that students were given the SET after they had completed the course and received their final course grade, which creates a situation that grade knowledge is officially available to students. This argument is also suggested by Cho and Cho (2017) who claimed that students adjust grade bias in SETs depending on how accurate the grade knowledge they have is.

The findings of the research were also in line with the discussion presented by Feldman (1976). Feldman underlined that the association between higher grades and higher SET scores may be causally a result of interest in the course or motivation. Feldman reviews the studies whose unit of analysis is individual students, and claims that students who took a course as an elective expressed higher approval when compared to the students who took the course as a requirement. A similar pattern also exists in this study. Although the research is not designed to ask students if they took the course as an elective or as a requirement, it is known that "S" grade is given in the non-credit courses which are taken as electives. Although the sample size which only consists of 4 students is very small and this is a limitation, it should be noted that the highest mean score for the SET instrument came from the students who stated that they got an "S" grade (M=116,75).

Furthermore, in this study, the descriptive mean statistics reveal the pattern that the higher the students' grades are, the higher the student evaluation of teaching effectiveness scores are. This pattern was also seen in Eiszler (2002). In Eiszler's study, it was found that faculty members received higher scores of SETs in the semesters when they distributed higher percentages of A and A- grades.

It is also important to evaluate these findings in the context of Turkish Education System as our participants come from this system. In this context, the significant issues to be discussed should be privatization of public education, student perceptions as customers, a focus on grades; and as a result, a system which puts the emphasis on grades rather than learning. Firstly, Kartal (2008) claims that there has been an increase in the number of private schools in Turkey since 1980s, and this increase can be said to be in parallel to the globalization trend in the world. Kartal adds that the state supports these schools in different ways with tax reductions and financial aids. He criticizes this situation because seeing students as the source of money contradicts with the real purpose of education. In other words, what can be concluded is that students become customers who need to be satisfied. Secondly, apart from the increase in the number of private schools and growing trend to see students as customers, national examination systems have brought other concerns like inflated grades. Dinç and Akşit (2015) examined the newspaper articles regarding concerns about the national examinations regarding entrance into high school. The grades student get during their primary school education has a contribution to this national exam as Primary School Education Success Grade (İlköğretim Başarı Puanı). Therefore, one concern coming from this contribution is inflated grades according to analysis of newspaper articles as mentioned in Dinç and Akşit (2015).

Therefore, when public education privatized, schools are competing to attract more students; and as a result; they will want to become more successful in the national examinations for high school and university entrance. As students' school grades has contributions to these examination results, all these create a specific focus on grade in Turkish education system. Coming from such a grade-focused education system, students may have tendency put grades before learning, and may evaluate their instructor not depending on how they teach but what grade they receive from these instructors. These are the concerns about grades in Turkish education system although it is important to warn the reader that this does not mean that private schools inflate their grades to satisfy the needs of the students.

As a result, when all these are taken into consideration, administrators and faculty need to be careful and aware of grade bias while using student evaluation of teaching effectiveness. As mentioned earlier, Eiszler (2002) found that even when the variables such as the influences of previous success, course attractiveness and charm of certain instructors were controlled, faculty received higher scores of SETs in the semesters when they distributed higher percentages of A and A- grades. This study was conducted to investigate if student evaluations of teaching effectiveness contributed to a tendency to inflate grades at a mid-sized public university in the midwestern United States. The data come from 983,491 teaching evaluations from students who were registered in 37,000 course sections during the spring 1980 and the fall of 1999. Eiszler also concludes that the trends in the elevation of expected grades in the current study at least partly a proof of inflated grades under the impact of student evaluations of teaching. Therefore, the author warns us that while the validity of SETs as a measurement of teaching effectiveness were proved, it is still possible to use them in ways that can cause gradual grade inflation. Therefore, administrators and faculty need to find ways to use SETs without causing grade inflation.

Firstly, one suggestion to handle grade inflation issue comes from Cho and Cho (2017). As the quality of grade knowledge affects how students evaluate their instructors, Cho and Cho suggests that instead of giving SETs after the final exam, which may allow students to either revenge or reward faculty, the implementation of SETs may be designed as one in the middle and one at the end of the academic year.

Secondly, Stroebe (2016) discusses some policy implications regarding grade inflation and student evaluations of teaching. Accordingly, SET concerns make faculty teach in ways that will require little work but higher grades with the expectation that they will end in higher ratings. The policy suggestions made by Stroebe include statistical adjustments in data analysis that will remove bias, using additional information like teaching portfolios to evaluate faculty performance. For the administrators, Stroebe discusses the necessity of deemphasizing the importance

of SET scores especially for the younger faculty members who feel under more pressure. Moreover, he suggests abandoning the fashion of honoring top teachers because getting the top results in SETs is not a guarantee of teaching quality.

Thirdly, Schneider (2013) concludes that better student learning requires a stricter and more challenging instruction, but many faculty members hold the belief that this type of instruction will damage their SET scores; and as a result, stand as an obstacle to getting tenure and promotion. Therefore, Schneider suggests that a reform in SETs is necessary in that they include items related to how much students have learnt and how much work they have put in the course.

Lastly, instead of relying only one aspect of faculty evaluation, administrators can also use peer reviews of teaching and teaching portfolios which are explained in Chapter 2. Since all techniques of evaluation have their own advantages and disadvantages, being aware of these advantages and disadvantages will provide administrators with opportunities to combine them in the best possible way.

The findings of the study also showed that the students of Humanities scored their instructor's teaching effectiveness significantly higher when compared to the students of Architecture and Natural Sciences. This finding is in line with the literature on teaching effectiveness, which claims that there is not a single valid definition of teaching effectiveness (Ryans, 1949; Arreola, 2007). The research on teaching effectiveness shows us that teaching is not a one-way linear act directed from teachers to students, but it is rather "a shared process" (Bidabadi, Isfahani, Rouhollahi, & Khalili, 2016) between teachers and students. Therefore, students' roles and expectations in this process should not be ignored. This is also in line with Allan, Clarke and Jopling (2009) who highlight student expectations in the evaluations of teaching and claim that students themselves have a very important role in making teaching effective.

As is clear from the previous paragraph, student expectations are important in the evaluations of teaching effectiveness. This brings the question of according to what these expectations change. As an answer to this question, context becomes vital. The importance of teaching context in the evaluations of teaching effectiveness has also been discussed in the literature. As the study by Parpala, Lindblom Ylänne and Rytkönen (2011) shows students in different disciplines may conceptualize teaching effectiveness in different ways. As also mentioned by Healey (2000) teaching quality is not independent of context and it has more than one dimension. Therefore, Healey underlines the importance of having diverse scholarships of teaching across different disciplines. Similarly, Brown and Atkins (1988) suggested that teaching effectiveness is affected by one's goals and a teaching practice which is effective in one context may not be effective in a different context. Devlin and Samarawickrema (2010) also argue that the definition of effective teaching transforms in response to the changes in the context where teaching and learning occur. All in all, as each discipline has a different teaching context, it is important to recognize the requirements of each context to make teaching effective.

Student approaches to teaching and learning which are also directly related to the definition of teaching effectiveness may change depending on students' discipline as shown by Cashin and Downey (1995) and Parpala et al. (2010). Parpala et al. highlighted that there were significant differences in the way students approach learning by their disciplines and Cashin and Downey (1995) showed that there were significant differences between "what is taught", "student perceptions of what they learn" and "student perceptions of how they are taught" across disciplines. Therefore, students in this study may also have perceived teaching and learning differently by their disciplines. According to Cashin and Downey, one reason for this could be the course objectives in that different courses and course objectives require different teaching methods, and it makes sense that students give higher ratings to those teaching applications which fit their goals. These studies also underline the importance of student perceptions of their own learning, independent of real learning

that takes place. Therefore, students' perceptions of their own learning in an offered course may also have its reflections in the student evaluations.

Furthermore, teaching methods and styles that students are exposed to in their own disciplines most of the time may be shaping their perceptions of teaching effectiveness since studies by Norton et al. (2005) and Lindblom-Ylänne et al. (2006) show us that teachers' approaches to teaching also change depending on their disciplines. Norton et al. (2005) found that teachers' intentions and beliefs related to teaching change by their disciplines, and how teachers conceptualize teaching and their teaching context together shape these intentions. Similarly, Lindblom Ylänne et al. (2006) found systematic differences in teachers' approaches to teaching both across disciplines and across various teaching contexts. What these studies may mean as to the results in the present study is that the way faculty members across different disciplines conceptualize teaching effectiveness together with their teaching intentions and beliefs may also shape student expectations as to effective teaching. Students who are used to certain teaching methods in certain disciplines may be looking for similar conduct in other courses they take. However, the limitation in this study is that we do not know the disciplines of faculty members who were rated by students. Therefore, we cannot make any comments about the different conceptualizations of teaching effectiveness across different disciplines. Still, it seems plausible to make comments about the impacts of these common teaching methods on student expectations as we have the knowledge of students' discipline in the study. For example, for a student who is used to observe teaching in a student-focused fashion in one department may experience difficulties when he/she takes an elective course which is offered in a teacher-focused fashion in a different department.

In addition to these, student evaluation of teaching effectiveness instrument used in this study does not include any discipline-specific questions related to the characteristics of teaching and learning across disciplines and this could be one reason for discipline-related bias that the study has found. Finding reliability concerns

in decision making across disciplines about student evaluations of teaching, Barnes and Barnes (1993) warns us that using a standard instrument for all academic disciplines may not be appropriate when these concerns are taken into consideration. However, the real implementation at Middle East Technical University is also done through a single standardized instrument; therefore, these findings are valuable in revealing possible bias in the implementation of student evaluation systems at the university.

Lastly, Cashin (1990) also aims to present possible explanations for disciplinary differences in student evaluations of teaching. According to Cashin, one possible explanation for the fact that more quantitative courses get lower scores in student evaluations is that about students' less-developed quantitative skills in comparison to their verbal skills. This could make teaching quantitative courses more difficult as students have lower expectations of achievement. Cashin also states that the reason for less specific courses to obtain higher ratings may be because there is possibly more than one correct answer for the questions, which brings more freedom and more opportunities to become successful; and this may be reflected in teaching evaluations by students. These disciplinary differences may be due to the differences in students' attitudes, academic talents and aims, motivation, learning styles and approaches to effective teaching.

5.2 Implications

The findings of the research when taken into consideration with the existing research on the topic have some implications for administrators, faculty members and students.

5.2.1 Implications for administrators

As we have seen, student evaluations of teaching effectiveness can be used both for summative and formative purposes by the administrators. Therefore, if there are any

bias in the results of these instruments, some revisions may be necessary to use them for decision-making purposes for the faculty. If used for any reasons, administrators need to be aware of their limitations. This is one of the most important reasons why the findings of this study have some implications for the administrators.

Administrators should train both faculty members and students about the possible grade bias in student evaluations of teaching. As SETs are good sources of information and feedback for the faculty to improve their teaching, faculty and administrators can work together to write the SET items that assess teaching effectiveness according to the needs of the institution. Faculty should be trained about what makes effective teaching based on these needs. As to the students, they should be trained about the emotional aspects of filling in SETs, they should be trained about the real purpose of SETs which is to improve teaching effectiveness, and students should know that using them as a reward to or revenge against their instructors due to their final course grade will not do any good to improve instructor's real teaching effectiveness.

As the accuracy of students' grade knowledge have impacts on how students evaluate their instructors, administrators may design the implementation of SETs as one in the middle of the term and one at the end of the term. They should make sure that the first one takes place before students have any knowledge of grade, and one takes place at the end of the academic term. Administrators should also ask for the grade or expected grade knowledge while designing these instruments so that they can include this information in their analysis, and see if there is any grade bias in the results of student evaluations of teaching effectiveness instrument.

As the findings of this study reveal, receiving good scores in SETs may not always be a guarantee of good teaching. Therefore, the administers should abandon honoring top teachers simply and only depending on SET scores. Instead, a combination of peer reviews of teaching, teaching portfolios may be included in the evaluation process besides SETs. This would also prevent concerns like grade inflation because

students are not the only source of information in the evaluation of teaching effectiveness.

As mentioned earlier, student learning requires more challenging instruction. Therefore, if administrators want SETs to assess teaching quality, they should include items that assess how much students have learnt and how much work and effort they have put in the course. Otherwise, the good scores in SETs may be a result of light workload and inflated grades.

In addition, administrators should be aware of these possible biases and as this study shows evaluations given to elective courses may be different than the evaluations given to must courses. What this means for administrators may be not comparing the scores of elective courses to the scores of must courses as if they occur on an equal basis. Other examples include comparing courses with different characteristics like student number and course difficulty. When the findings of the third research question are also taken into consideration, administrators should also recognize the requirements of each teaching context and interpret the differences in the results of student evaluations of teaching accordingly especially if the results are being used for summative purposes.

5.2.2 Implications for faculty members and students

The findings of this study have also some implications for faculty members on the grounds that SETs have formative uses to improve their teaching quality. Faculty should be aware of grade bias in the interpretation of their SET scores so that they can intervene in and revise the practices when necessary. They can also ask for continuous verbal and written feedback from their students at different points in the term to improve their teaching. In addition, faculty members should not forget that effective teaching does not only have one definition, and it can depend on fulfilment of the requirements of the teaching context. Therefore, faculty members should be eager to adapt their teaching according to the needs of the context and their students.

When research on peer review teaching and teaching portfolios are also considered, faculty members should work on their own teaching philosophy and develop self-awareness. They can also sometimes ask for feedback about their teaching and course materials not only from students but also from their colleagues. Faculty collaboration to improve teaching becomes significant when the biases in student evaluations of teaching are considered.

As low-response rates decrease the reliability of the scores, faculty members should inform their students about the importance of filling in the student evaluation forms, should encourage students to give honest answers to the questions to improve the teaching quality.

Faculty members should take into consideration the feedback they receive from students, but as the literature review shows, they should also know that students are not fully knowledgeable to evaluate all aspects of teaching perfectly. They should also make adaptations to their teaching with this awareness.

In addition, if their university have a student evaluation instrument which is prone to the biases mentioned and discussed in this study, faculty members should not simply rely on this instrument to enhance their teaching. They can determine the points they want to improve and directly ask feedback from their students.

5.3 Recommendations for Future Research

Based on the findings of the present study, the following recommendations can be made for future research:

- A mix-method study can be designed to have a deeper understanding of how the scores in student evaluations of teaching effectiveness differ by student's grade and discipline, and how students perceive and interpret teaching

effectiveness of male and female faculty members. To this end, interviews can be conducted with both students and faculty members.

- The intersection of faculty member's gender, age, title, and managerial roles together can be studied to have a broader understanding of how students perceive faculty member's teaching effectiveness depending on these variables.
- Qualitative research design can be used and interviews can be conducted to have a better understanding of what teaching effectiveness means for students from different disciplines.

REFERENCES

Acar, F. (1996). Türkiye'de kadın akademisyenler: Tarihsel evrim ve bugünkü durum. In H. Coşkun (Ed.), *Akademik yaşamda kadın: Türk ve Alman üniversitelerinde kadın kariyerlerinin karşılaştırılması* (pp.75-88). Ankara: Türk-Alman İşleri Kurulu.

Acker, J. (1990). Hierarchies, jobs, bodies: A theory of gendered organizations, *4*(2), 139–158.

Acker, J. (1992). From sex roles to gendered institutions. *Contemporary Sociology*, *21*(5), 565-569.

Acker, J. (2009). From glass ceiling to inequality regimes. *Sociologie Du Travail*, *51*(2), 199–217. doi: 10.1016/j.soctra.2009.03.004

Akbulut, N. (2011). Ayna benlik: Kadın akademisyen gelenekçidir! *Eğitim Bilim Toplum Dergisi*, *9*(34), 50-66.

AL-Saghir, M. (2008). The validity of students' evaluations of teaching effectiveness (SETS). *Association for University Regional Campuses of Ohio*, *14*, 7-18.

Alemu, B. M. (2014). Enhancing the quality and relevance of higher education through effective teaching practices and instructors' characteristics. *Universal Journal of Educational Research*, *2*(9), 632–647. doi: 10.13189/ujer.2014.020906

Alkoç, Y. (2017). Üniversite imajı, üniversiteye duyulan memnuniyet ve öğrenci sadakati arasındaki ilişkileri anlamaya yönelik bir araştırma. *Uluslararası Bilimsel Araştırmalar Dergisi*, *2*(2), 270–280. doi: 10.21733/ibad.2135

Allan, J., Clarke, K., & Jopling, M. (2009). Effective teaching in higher education: Perceptions of first year undergraduate students. *International Journal of Teaching and Learning in Higher Education*, *21*(3), 362–372.

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

Arnăutu, E., & Panc, I. (2015). Evaluation criteria for performance appraisal of faculty members. *Procedia - Social and Behavioral Sciences*, *203*, 386–392. doi:10.1016/j.sbspro.2015.08.313

Arreola, R. A. (2007). *Developing a comprehensive faculty evaluation system*. USA: Jossey-Bass.

Ashmore, R. D., & Del Boca, F. K. (1979). Sex stereotypes and implicit personality theory: toward a cognitive-social psychological conceptualization, *Sex Roles*, *5*(2), 219-248.

Bailyn, L. (2003). Academic careers and gender equity: Lessons learned from MIT. *Gender, Work and Organization*, *10*(2), 137–153. doi:10.1111/1468-0432.00008

Baker P., & Copp, M. (1997). Gender matters most: The interaction of gendered expectations, feminist course content , and pregnancy in student course evaluations, *American Sociological Association*, *25*(1), 29–43.

Balyer, A., & Gündüz, Y. (2011). Türk yükseköğretim yönetim sisteminde YÖK ile yaşanan paradigmatik dönüşüm: Vakıf üniversiteleri çelişkisi. *Sosyal Bilimler Enstitüsü Dergisi*, *31*(2), 69–84. Retrieved from http://dergipark.gov.tr/download/article-file/219585

Barnard, A., Nash, R., McEvoy, K., Shannon, S., Waters, C., Rochester, S., & Bolt, S. (2015). LeaD-In: a cultural change model for peer review of teaching in higher education. *Higher Education Research and Development*, *34*(1), 30–44. doi: 10.1080/07294360.2014.935931

Barnes, L. L. B., & Barnes, M. W. (1993). Academic discipline and generalizability of student evaluations of instruction. *Research in Higher Education*, *34*(2), 135-149. doi: 10.1007/BF00992160

Basow, S. A. (1992). *Gender stereotypes and roles*. US: Brooks/Cole Publishing Company.

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

Başarır, F., & Sarı, M. (2015). Kadın akademisyenlerin "kadın akademisyen olma"ya ilişkin algılarının metaforlar yoluyla incelenmesi. *Yükseköğretim ve Bilim Dergisi, 5*(1), 41-51. doi: 10.5961/jhes.2015.108

Baykal, Ü., Sökmen, S., & Korkmaz Akgün, E. (2002). Öğrenci memnuniyeti ölçeği geliştirme çalışması. *Hemşirelik Dergisi, 11*(49), 23-32.

Berk, R. A. (2006). *Thirteen strategies to measure college teaching*. USA: Stylus Publishing.

Bidabadi, N. S., Isfahani, A. N., Rouhollahi, A., & Khalili, R. (2016). Effective teaching methods in higher education: Requirements and barriers. *Journal of Advances in Medical Education & Professionalism, 4*(4), 170–178.

Bird, S. R. (2011). Unsettling universities' incongruous, gendered bureaucratic structures: A case-study approach. *Gender, Work & Organization, 18*, 202-230. doi:10.1111/j.1468-0432.2009.00510.x

Bobbitt-Zeher, D. (2011). Gender discrimination at work: Connecting gender stereotypes, institutional policies, and gender composition of workplace. *Gender and Society, 25*(6), 764–786. doi:10.1177/0891243211424741

Borich, G. D. (2007). *Effective teaching methods research-based practice*. USA: Pearson.

Boring, A. (2016). Gender biases in student evaluations of teaching. *Journal of Public Economics, 145*, 27–41. doi: 10.1016/j.jpubeco.2016.11.006

Boring, A., Ottoboni, K., & Stark, P. B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*, 1-11. doi: 10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

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

Brown, G., & Atkins, M. (1988). *Effective teaching in higher education*. USA: Routledge.

Burton, C. (1991). Job evaluation and gender bias. *Journal of Tertiary Education Administration, 13*(1), 73–81. doi: 10.1080/0157603910130107

Carli, L. L., Alawa, L., Lee, Y. A., Zhao, B., & Kim, E. (2016). Stereotypes about gender and science: Women ≠ scientists. *Psychology of Women Quarterly, 40*(2), 244–260. doi: 10.1177/0361684315622645

Cashin, W. E. (1990). Students do rate different academic fields differently. *New Directions for Teaching and Learning*, 113-121. doi:10.1002/tl.37219904310

Cashin, W. E., & Downey, R. G. (1995). Disciplinary differences in what is taught and in students' perceptions of what they learn and of how they are taught. *New Directions for Teaching and Learning*, 81-92. doi:10.1002/tl.37219956412

Centra, J., & Gaubatz, N. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education, 71*(1), 17-33. doi:10.2307/2649280

Chalmers, D. (2011). Student feedback in the Australian national and university context. In C. S. Nair, & P. Mertova (Eds.), *Student feedback: The cornerstone to an effective quality assurance system in higher education* (pp. 81-97). UK: Chandos Publishing.

Channing, J. (2016). Faculty evaluations: Contentious bothers or important tools for faculty growth? *Community College Journal of Research and Practice, 41*(11), 757–760. doi: 10.1080/10668926.2016.1241197

Chism, N. V. N. (2007). *Peer review of teaching*. USA: Anker Publishing Company.

Cho, D., Baek, W., & Cho, J. (2015). Why do good performing students highly rate their instructors? Evidence from a natural experiment. *Economics of Education Review, 49*, 172-179. doi: 10.1016/j.econedurev.2015.10.001

Cho, D., & Cho, J. (2017). Does more accurate knowledge of course grade impact teaching evaluation? *Education Finance and Policy, 12*(2), 54. doi:10.1162/EDFP_a_00197

Christiaens, W., Spooren, P., Mortelmans, D., Van Loon, F. J. A. (2014). Students' perceptions of learning, course grades, and student evaluation of teaching: An empirical analysis. *The International Journal of Assessment and Evaluation*, *20*(3), 13-21.

Chung, Y. B., Marshall, J. A., & Gordon, L. L. (2001). Multiculturally-focused counseling supervision: Its relationship to trainees' multicultural counseling self-efficacy. *The Clinical Supervisor, 20*(1), 87–98. doi:10.1300/J001v20n01

Curry, E. (2006). Faculty performance reviews. *Tomorrow's Professor*, *1*(2), 1–3.

Danyal-Köker, E. (1988). *Türkiye'de kadın, eğitim ve siyaset "Yüksek öğrenim kurumlarında kadının durumu üzerine bir inceleme"* (Doctoral thesis, Ankara University, Ankara, Turkey).

Das, M., & Das, H. (2001). Business students' perceptions of best university professors: Does gender role matter? *Sex Roles, 45*(9/10), 665–676.

Devanas, M. A. (2006). Teaching portfolios. In P. Seldin (Ed.), *Evaluating faculty performance* (pp. 111-130). USA: Jossey-Bass.

Devlin, M., & Samarawickrema, G. (2010). The criteria of effective teaching in a changing higher education context. *Higher Education Research and Development*, *29*(2), 111–124. doi: 10.1080/07294360903244398

Dinç, E. , & Akşit, İ. (2015). Ortaöğretime geçiş sınavları ile ilgili köşe yazılarının incelenmesi. *Türkiye Sosyal Araştırmalar Dergisi*, *2*(2), 59-82. Retrieved from http://dergipark.gov.tr/tsadergisi/issue/21493/230408

Donaghue, H., & Dolci, I. (2013). Professional teaching portfolios. *TESOL Arabia Perspectives, 21* (2), 13-17. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=plh&AN=482437&site=bsi-live

Eiszler, C. F. (2002). College students' evaluations of teaching and grade inflation. *Research in Higher Education*, *43*(4), 483-501.

Elez, N. (2017). Acceptance of marketing concept at higher education institutions-student focused approach. *Zbornik Radova Ekonomskog Fakulteta u Istocnom Sarajevu*, *1*(14), 61-70. doi: 10.7251/ZREFIS1714061E

Emil, S. (2013, September). Pilot study on the students' perceptions of effective university teachers in a Turkish university. A paper presented at the European Conference on Educational Research, İstanbul, Turkey.

Erginer, A., & Saklan, E. (2016). The problematic of transition of teaching profession to a woman's occupation. *Uluslararası Toplum Araştırmaları Dergisi*, *6*(11), 415-450.

Erikson, M. G., Erlandson, P., & Erikson, M. (2015). Academic misconduct in teaching portfolios. *International Journal for Academic Development*, *20*(4), 345–354. doi: 10.1080/1360144X.2015.1083435

Feldman, K. A. (1976). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education*, *4*(1), 69–111.

Feldman, K.A. (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

Fernández, J., & Mateo, M. A. (1997). Student and faculty gender in ratings of university teaching quality. *Sex Roles*, *37*(11/12), 997-1003. doi: 10.1007/BF02936351

Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. UK: Sage.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education*. USA: McGraw-Hill.

Gravetter, F. J., & Wallnau, L. B. (2013). *Statistics for the behavioral sciences*. CANADA: Cengage Learning.

Hamilton, M. (2017). Bridging the gap from teacher to teacher educator: The role of a teaching portfolio. *Studying Teacher Education*, *14*(1), 88–102. doi:10.1080/17425964.2017.1414041

Harvey, L. (2011). The nexus of feedback and improvement. In C. S. Nair, & P. Mertova (Eds.), *Student feedback: The cornerstone to an effective quality assurance system in higher education* (pp. 3-28). UK: Chandos Publishing.

Healey, M. (2000). Developing the scholarship of teaching in higher education: A discipline-based approach. *Higher Education Research & Development*, *19*(2), 169-189. doi: 10.1080/072943600445637

Hénard, F., & Roseveare, D. (2012). *Fostering quality teaching in higher education: Policies and practices*. Retrieved from https://www.oecd.org/education/imhe/QT%20policies%20and%20practices.pdf

Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of Educational Psychology*, *63*(2), 130-133. doi: 10.1037/h0032636

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education*, *4*(1), 1-8. doi: 10.1080/2331186X.2017.1304016

Isely, P., & Singh, H. (2005). Do higher grades lead to favorable student evaluations? *The Journal of Economic Education*, *36*(1), 29-42. Retrieved from http://www.jstor.org/stable/30042630

Joshi, A. (2014). By whom and when is women's expertise recognized? The interactive effects of gender and education in science and engineering teams. *Administrative Science Quarterly*, *59*(2), 202–239. doi: 10.1177/0001839214528331

Joye, S. W., & Wilson, J. H. (2015). Professor age and gender affect student perceptions and grades. *Journal of the Scholarship of Teaching and Learning*, *15*(4), 126–138. doi: 10.14434/josotl.v15i4.13466

Kartal, S. (2008). Türk eğitim sisteminde özel okullar. *Mülkiye Dergisi*, *32*(258), 135-150.

Kalaycı, N. (2009). Yükseköğretim kurumlarında akademisyenlerin öğretim performansını değerlendirme sürecinde kullanılan yöntemler. *Kuram ve Uygulamada Eğitim Yönetimi*, *15*(60), 625–656. Retrieved from http://dergipark.ulakbim.gov.tr/kuey/article/view/5000050585

Keig, L., & Waggoner, M. D. (1995). *Collaborative peer review. The role of faculty in improving college teaching*. USA: ERIC Clearinghouse on Higher Education.

Kember, D., & Ginns, P. (2012). *Evaluating teaching and learning: A practical handbook for colleges, universities and the scholarship of teaching*. USA: Routledge.

Kerchner, C. T., & Mitchell, D. E. (1986). Teaching reform and union reform. *The Elementary School Journal*, *86*(4), 448–470.

Lavigne, A. L., & Good, T. (2014). *Teacher and student evaluation: moving beyond the failure of school reform*. USA: Routledge.

Lester, J. (2008). Performing gender in the workplace. *Community College Review*, *35*(4), 277–305. doi:10.1177/0091552108314756

Lin, T. (2015). A closer look at the relationship between grades and teacher evaluation: The role of grade knowledge. *The Journal of Economics*, *41*(2), 35–66.

Lin, K. J., Chen, H. M., Chen, H.M., & Chang, Y.S. (2017). The effect of TQM strategy on learning satisfaction and loyalty of students - The mediation effect of teaching quality. *World Transactions on Engineering and Technology Education*, *15*(1), 49–56.

Lindblom-Ylänne, S., Trigwell, K., Nevgi, A., & Ashwin, P. (2006). How approaches to teaching are affected by discipline and teaching context. *Studies in Higher Education*, *31*(3), 285-298. doi: 10.1080/03075070600680539

Lorber, J. (1994). *Paradoxes of gender*. USA: Yale University Press.

Love, D. A., & Kotchen, M. J. (2010). Grades, course evaluations, and academic incentives. *Eastern Economic Journal*, *36*(2), 151-163. doi: 10.1057/eej.2009.6

Lowder, L., Atiqulla, M., Colebeck, D., Das, S., Karim, M. A., Khalid, A., & Singh, R. (2017). Peer observation: Improvement of teaching effectiveness through class participation at a polytechnic university. *Journal of STEM Education*, *18*(4), 51–57.

MacNell, L., Driscoll, A., & Hunt, A. N. (2014). 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

Matos-Díaz, H., & Ragan Jr. J. F. (2010). Do student evaluations of teaching depend on the distribution of expected grade? *Education Economics*, *18*(3), 317-330, doi: 10.1080/09645290903109444

McKeachie, W. (1996). Student ratings of teaching (Occasional Paper No. 33). American Council of Learned Societies, University of Michigan. Retrieved from http://archives.acls.org/op/33_Professonal_Evaluation_of_Teaching.htm

Meltzer, A. L., & McNulty, J. K. (2011). Contrast effects of stereotypes: "Nurturing" male professors are evaluated more positively than "nurturing" female professors. *The Journal of Men's Studies*, *19*(1), 57–64. doi: 10.3149/jms.1901.57

Merriam Webster Learner's Dictionary. (n.d.). *"Equality"*. Retrieved on April 12, 2018 from http://learnersdictionary.com/definition/equality

Merriam Webster Learner's Dictionary. (n.d.). *"Equity"*. Retrieved on April 12, 2018 from http://learnersdictionary.com/definition/equity

Merriam Webster Learner's Dictionary. (n.d.). *"Stereotype"*. Retrieved on April 5, 2018 from http://learnersdictionary.com/definition/stereotype

METU (2012). *Student handbook*. Retrieved on September 10, 2018 from http://oidb.metu.edu.tr/sites/oidb.metu.edu.tr/files/ODTU%20Ogrenci%20El%20Kitabi.pdf

METU (2018). *General information*. Retrieved on September 10, 2018 from https://www.metu.edu.tr/tr/genel-bilgiler

METU Directorate of Personnel Department (2018). *Academic staff appointment and promotion criteria*. Retrieved on September 10, 2018 from http://pdb.metu.edu.tr/ogretim-uyesi-atama-ve-yukselme-kriterleri

METU Directorate of Student Affairs (2018). *Credit system of the courses*. Retrieved on September 10, 2018 from http://oidb.metu.edu.tr/derslerin-kredi-sistemi

METU Parlar Foundation (2018). *Award regulation*. Retrieved September 10, 2018 from http://parlar.org.tr/odul-yonetmeligi/

Miller, J., & Chamberlin, M. (2000). Women are teachers, men are professors: A study of student perceptions. *Teaching Sociology*, *28*(4), 283-298. Retrieved from http://www.jstor.org/stable/1318580

Mittal, S., Gera, R., & Batra, D. K. (2015). Evaluating the validity of student evaluation of teaching effectiveness (SET) in India. *Education + Training*, *57*(6), 623-638. doi: 10.1108/ET-06-2013-0072

Morley, D. (2014). Assessing the reliability of student evaluations of teaching: choosing the right coefficient. *Assessment & Evaluation in Higher Education*, *39*(2), 127–139. doi:10.1080/02602938.2013.796508

Nasser-Abu Alhija, F. (2016). Teaching in higher education: Good teaching through students' lens. *Studies in Educational Evaluation, 54,* 4–12. doi:10.1016/j.stueduc.2016.10.006

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

Nilson, L. B. (2010). *Teaching at its best: A research-based resource for college instructors*. USA: Jossey-Bass.

Norton, L., Richardson, T. E., Hartley, J., Newstead, S., & Mayes, J. (2005). Teachers' beliefs and intentions concerning teaching in higher education. *Higher Education*, *50*, 537-571. doi: 10.1007/s10734-004-6363-z

Nowell, C. (2007). The impact of relative grade expectations on student evaluation of teaching. *International Review of Economics Education*, *6*(2), 42-56. doi: 10.1016/S1477-3880(15)30104-3

Nussbaum, M. C. (2000). *Women and human development: The capabilities approach*. Cambridge: Cambridge University Press.

Onwuegbuzie, A.J., Daniel, L.G. & Collins, K.M.T. (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity*, *43*(2), 197-209. doi: 10.1007/s11135-007-9112-4

Onwuegbuzie, A., Witcher, A., Collins, K., Filer, J., Wiedmaier, C., & Moore, C. (2007). Students' perceptions of characteristics of effective college teachers: A validity study of a teaching evaluation form using a mixed-methods analysis. *American Educational Research Journal*, *44*(1), 113-160. Retrieved from http://www.jstor.org/stable/30069473

Özbilgin, M., & Healy, G. (2004). The gendered nature of career development of university professors: The case of Turkey. *Journal of Vocational Behavior*, *64*, 358–371. doi: 10.1016/j.jvb.2002.09.001

Özkanli, Ö. (2007). The situation of academic women in Turkey. *Eğitim ve Bilim*, *32*(144), 59-70.

Öztan, E., & Doğan, S.T. (2015). Akademinin cinsiyeti: Yıldız Teknik Üniversitesi örneği üzerinden üniversite ve toplumsal cinsiyet. *Çalışma ve Toplum*, *3*, 191-221.

Padró, F. F. (2011). Student feedback in the US and global contexts. In C. S. Nair, & P. Mertova (Eds.), *Student feedback: The cornerstone to an effective quality assurance system in higher education* (pp. 29-48). UK: Chandos Publishing.

Pallet, W. (2006). Uses and abuses of student ratings. In P. Seldin (Ed.), *Evaluating faculty performance* (pp. 50-65). USA: Jossey-Bass.

Parpala, A. , Lindblom-Ylänne, S. , Komulainen, E. , Litmanen, T. and Hirsto, L. (2010). Students' approaches to learning and their experiences of the teaching–learning environment in different disciplines. *British Journal of Educational Psychology*, *80*, 269-282. doi:10.1348/000709909X476946

Parpala, A., Lindblom-Ylänne, S., & Rytkönen, H. (2011). Students' conceptions of good teaching in three different disciplines. *Assessment & Evaluation in Higher Education*, *36*(5), 549-563. doi: 10.1080/02602930903541023

Passer, M. W., & Smith, R. E. (2008). *Psychology: The science of mind and behavior*. USA: McGraw-Hill.

Piihl, J., & Rasmussen, J. S. (2013). Course evaluation systems for open-ended quality enhancement. In C. Nygaard, N. Courtney, & P. Bartholomew (Eds.). *Quality enhancement of university teaching and learning* (pp. 181-199). UK: Libri Publishing.

Price, L., Svensson, I., Borell, J. & Richardson, J.T.E. (2017). The role of gender in students' ratings of teaching quality in computer science and environmental engineering. *IEEE Transactions on Education*, *60*(4), 281-287. doi: 10.1109/TE.2017.2696904

Punyanunt-Carter, N., & Carter, S. (2015). Students' gender bias in teaching evaluations. *Higher Learning Research Communications*, *5*(3), 28-37. doi: 10.18870/hlrc.v5i3.234

Puteh, M., & Habil, H. (2011). Student feedback in higher education: A Malaysian perspective. In C. S. Nair, & P. Mertova (Eds.), *Student feedback: The cornerstone to an effective quality assurance system in higher education* (pp. 49-60). UK: Chandos Publishing.

Reeves, M. (2010). *Women in business: Theory, case studies, and legal challenges*. New York: Routledge.

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.

Ridgeway, C. L. (2009). Framed before we know it: How gender shapes social relations. *Gender and Society*, *23*(2), 145–160. doi:10.1177/0891243208330313

Rotenberg, R. (2005). *The art & craft of college teaching: A guide for new professors & graduate students*. USA: Left Coast Press.

Ryans, D. G. (1949). The criteria of teaching effectiveness. *The Journal of Educational Research*, *42*(9), 690–699.

Sandberg, S. (2013). *Lean in: Women, work and the will to lead*. USA: Alfred A. Knoff.

Schneider, G. (2013). Student evaluations, grade inflation and pluralistic teaching: Moving from customer satisfaction to student learning and critical thinking. *Forum for Social Economics*, *42*(1), 122–135. doi:10.1080/07360932.2013.771128

Seldin, P. (2006). Building a successful evaluation program. In P. Seldin (Ed.), *Evaluating faculty performance* (pp. 1-19). USA: Jossey-Bass.

Seldin, P., Miller, J.E., & Seldin, C. A. (2010). *The teaching portfolio: A practical guide to improved performance and promotion/tenure decisions*. USA: Jossey-Bass.

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

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

Stehle, S., Spinath, B., & Kadmon, M. (2012). Measuring teaching effectiveness: correspondence between students' evaluations of teaching and different measures of student learning. *Research in Higher Education, 53*(8), 888-904. doi: 10.1007/s11162-012-9260-9

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

Takiff, H. A., Sanchez, D. T., & Stewart, T. L. (2001). What's in a name? The status implications of students' terms of address for male and female professors. *Psychology of Women Quarterly*, *25*(2), 134–144. doi: 10.1111/1471-6402.00015

Tata, J. (1999). Grade distributions, grading procedures, and students' evaluations of instructors: A justice perspective. *The Journal of Psychology*, *133*(3), 263-271. doi: 10.1080/00223989909599739

The World Bank (2018). *Tertiary education, academic staff (% female) Turkey.* Retrieved on January 20, 2018 from: https://data.worldbank.org/indicator/SE.TER.TCHR.FE.ZS?locations=TR

Tindall, N. T. J., & Waters, R. D. (2017). Does gender and professional experience influence students' perceptions of professors? *Journalism & Mass Communication Educator*, *72*(1), 52–67. doi:10.1177/1077695815613932

UNESCO (2017). *Women in Science*. Retrieved on January 20, 2018 from http://uis.unesco.org/sites/default/files/documents/fs43-women-in-science-2017-en.pdf

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

van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2018). Developing an instrument for teacher feedback: Using the Rasch Model to explore teachers' development of effective teaching strategies and behaviors. *The Journal of Experimental Education*, *86*(2), 247–264. doi: 10.1080/00220973.2016.1268086

Wagner, N., Rieger, M., & Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review*, *54*, 79–94. doi:10.1016/j.econedurev.2016.06.004

West, C., & Zimmerman, D. H. (1990). Doing gender. In J. Lorber, & S. A. Farrell (Eds.), *The social construction of gender* (pp. 13-37). USA: Sage Publications.

Wheeless, V. E., & Potorti, P. F. (1989). Student assessment of teacher masculinity and femininity: A test of the sex role congruency hypothesis on student attitudes toward learning. *Journal of Educational Psychology, 81*(2), 259-262. doi: 10.1037/0022-0663.81.2.259

Wren, B. M. (2006). Examining gender differences in performance evaluations, rewards and punishments. *Journal of Management Research*, *6*(3), 115–124.

Yamawaki, N., Queiroz, A., & Hanson, P. (2012). The mediating role of expected grade on gendered teaching style biases in teacher evaluations. *Innovative Teaching*, *1*, 1-9. doi:10.2466/03.11.IT.1.1

Yiend, J., Weller, S., & Kinchin, I. (2014). Peer observation of teaching: The interaction between peer review and developmental models of practice. *Journal of Further and Higher Education*, *38*(4), 465–484. doi: 10.1080/0309877X.2012.726967

Young, S. & Shaw, D. G. (1999). Profiles of effective college and university teachers. *The Journal of Higher Education*, *70*(6), pp. 670-686, doi: 10.1080/00221546.1999.11780803

Yükseköğretim Kurulu (2018). *Öğretim elemanları sayısı özet tablosu* [Yükseköğretim Bilgi Yönetim Sistemi]. Retrieved from https://istatistik.yok.gov.tr/

111 devlet üniversitesinin sadece 3 tanesinin rektörü kadın. (2018, January). Habertürk. Retrieved from http://www.haberturk.com/111-devlet-universitesinin-sadece-3-tanesinin-rektoru-kadin-1788441

APPENDICES

A. Approval Letter from METU Human Subjects Ethics Committee

UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA/ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59
ueam@metu.edu.tr
www.ueam.metu.edu.tr

Sayı: 28620816/ 357

02 OCAK 2018

Konu: Değerlendirme Sonucu

Gönderen: ODTÜ İnsan Araştırmaları Etik Kurulu (İAEK)

İlgi: İnsan Araştırmaları Etik Kurulu Başvurusu

Sayın Yrd.Doç.Dr. Serap EMİL;

Danışmanlığını yaptığınız yüksek lisans öğrencisi Ece YILMAZ'ın " **Öğretim elemanlarının öğrenciler tarafından değerlendirmelerinde cinsiyet faktörü"** başlıklı araştırması İnsan Araştırmaları Etik Kurulu tarafından uygun görülerek gerekli onay **2017-EGT-197** protokol numarası ile **02.01.2018-28.09.2018** tarihleri arasında geçerli olmak üzere verilmiştir.

Bilgilerinize saygılarımla sunarım.

Prof. Dr. Ş. Halil TURAN
Başkan V

Prof. Dr. Ayhan SOL
Üye

Prof. Dr. Ayhan Gürbüz DEMİR
Üye

Doç. Dr. Yaşar KONDAKÇI
Üye

Doç. Dr. Zana ÇİTAK
Üye

Yrd. Doç. Dr. Pınar KAYGAN
Üye

Yrd. Doç. Dr. Emre SELÇUK
Üye

B. Consent Form for Data Collection

Bu çalışma, Orta Doğu Teknik Üniversitesi Eğitim Bilimleri Bölümü öğretim üyesi Y. Doç. Dr. Serap Emil gözetiminde, Eğitim Yönetimi ve Planlaması Yüksek Lisans Programı öğrencisi Ece Yılmaz tarafından yürütülmektedir. Bu form sizi araştırma koşulları hakkında bilgilendirmek için hazırlanmıştır.

Çalışmanın amacı öğrencilerin hangi özelliklerine dayanarak ders almış oldukları öğretim elemanlarının etkili veya etkisiz olduklarını düşündüklerinden yola çıkarak öğretim elemanlarının etkili veya etkisiz düşünülmesine sebep veren özellikleri tespit etmektir. Araştırmaya katılmayı kabul ederseniz, sizden beklenen hazırlamış olduğumuz anketi doldurarak bu konulardaki görüşlerinizi bize iletmenizdir. Soruları cevaplayabilmeniz için öncelikle şimdiye kadar üniversitede ders almış olduğunuz öğretim elemanlarından öğretimde etkili veya etkisiz olduğunu düşündüğünüz birini seçmeniz gerekmektedir. Araştırma şu an kaçıncı sınıfta olduğunuz, bölümünüz, seçmiş olduğunuz öğretim elemanının cinsiyeti, sizin cinsiyetiniz, genel not ortalamanız ve seçtiğiniz öğretim elemanının dersinden aldığınız not ile ilgili sorular içeren bir bölümle başlamaktadır. Bu soruları, öğretim elemanını değerlendireceğiniz bir öğretim elemanı değerlendirme anketi takip edecektir. Çalışma, uzunlukları birbirinden farklı 2 bölümden oluşmakta ve yaklaşık 10 dakika sürmektedir.

Araştırmaya katılımınız tamamen gönüllülük temelinde olmalıdır. Ankette, sizden kimlik veya kurum belirleyici hiçbir bilgi istenmemektedir. Cevaplarınız tamamıyla gizli tutulacak, sadece araştırmacılar tarafından değerlendirilecektir. Katılımcılardan elde edilecek bilgiler toplu halde değerlendirilecek ve bilimsel yayımlarda kullanılacaktır. Sağladığımız veriler gönüllü katılım formlarında toplanan kimlik bilgileri ile eşleştirilmeyecektir. Anket genel olarak kişisel rahatsızlık verecek sorular içermemektedir. Ancak katılım sırasında sorulardan ya da herhangi başka nedenden ötürü kendinizi rahatsız hissederseniz cevaplama işini yarıda bırakıp araştırmadan ayrılabilirsiniz.

Anket sonunda bu çalışmayla ilgili sorularınız cevaplanacaktır. Araştırmaya katıldığınız için şimdiden teşekkür ederiz. Çalışma hakkında daha fazla bilgi almak için Eğitim Yönetimi ve Planlaması Bölümü yüksek lisans öğrencisi Ece Yılmaz (e-posta: yilmaz.ece@metu.edu.tr) ya da Eğitim Bilimleri Bölümü öğretim üyelerinden Y. Doç. Dr. Serap Emil (e-posta: semil@metu.edu.tr) ile iletişim kurabilirsiniz.

Yukarıdaki bilgileri okudum ve bu çalışmaya tamamen gönüllü olarak katılıyorum. (Formu imzaladıktan sonra uygulayıcıya geri veriniz.)
Tarih:
Ad/Soyad:
İmza:

C. Instrument

Öğretim Elemanı Değerlendirme Anketi

Öğretim Elemanı Değerlendirme Anketi'ni cevaplamadan önce lütfen üniversite yaşantınız boyunca ders almış olduğunuz bir öğretim elemanını belirleyiniz. Bu öğretim elemanı, eğitim öğretim elemanı, eğitim-öğretim açısından etkili ya da etkisiz olduğunu düşündüğünüz bir kişi olmalıdır.

Bu kişiyi mümkünse yakından tanıyor ve ders aldığınız diğer öğretim elemanları ile karşılaştırabiliyor olmanız anketin güvenirliği açısından önemlidir. Seçtiğiniz öğretim elemanının **dönem boyunca göstermiş olduğu performansı göz önünde bulundurarak**, ankette sunulan her bir maddeyi, 1'den 5'e kadar verilen ölçme kriterlerine göre değerlendiriniz. Eğer herhangi bir maddeyle ilgili olarak fikriniz yok ise 0 (Uygun Değil) seçeneğini işaretleyiniz.

Bu ankete katıldığınız için çok teşekkür ederiz.

KİŞİSEL BİLGİLER

1. Şu an kaçıncı sınıftasınız? _____
2. Bölümünüz: _____
3. Seçmiş olduğunuz öğretim elemanının cinsiyeti: _____
4. Sizin cinsiyetiniz: _____
5. Genel not ortalamanız: _____
6. Seçtiğiniz öğretim elemanının dersinden aldığınız not: _____

ÖĞRETİM ELEMANI DEĞERLENDİRME ANKETİ

1: Kesinlikle katılmıyorum	3: Kararsızım	5: Kesinlikle katılıyorum
2: Katılmıyorum	4: Katılıyorum	0: Uygun değil

Seçmiş olduğunuz öğretim elemanına dair cevaplanması beklenen maddeler aşağıdadır.	Kesinlikle katılmıyorum	Katılmıyorum	Kararsızım	Katılıyorum	Kesinlikle katılıyorum	Uygun değil
1. Öğretim elemanı kendi alanı hakkında bilgiliydi.	1	2	3	4	5	0
2. Öğretim elemanı etkili iletişim kurardı.	1	2	3	4	5	0
3. Öğretim elemanı öğretim konusunda hevesliydi.	1	2	3	4	5	0
4. Öğretim elemanı işlenen her derse hazırlıklı gelirdi.	1	2	3	4	5	0
5. Öğretim elemanı rahat bir öğrenme atmosferi yaratmıştı.	1	2	3	4	5	0
6. Öğretim elemanı öğrenci ihtiyaçlarını göz önünde bulundururdu.	1	2	3	4	5	0
7. Öğretim elemanı farklı görüş ve fikirlere karşı hoşgörülüydü.	1	2	3	4	5	0
8. Öğretim elemanı öğrencilere karşı saygılı davranırdı.	1	2	3	4	5	0
9. Öğretim elemanı sıcak ve arkadaşça davranırdı.	1	2	3	4	5	0
10. Öğretim elemanı iyi bir mizah anlayışına sahipti.	1	2	3	4	5	0
11. Öğretim elemanı öğrencilerin yapabileceklerinin en iyisini yapmaları için motive ederdi.	1	2	3	4	5	0
12. Öğretim elemanının kendine güveni vardı.	1	2	3	4	5	0
13. Öğretim elemanı öğretmekten keyif alırdı.	1	2	3	4	5	0
14. Öğretim elemanı öğrencilerin öğrenmesini dikkate alırdı.	1	2	3	4	5	0
15. Öğretim elemanı materyalleri net bir şekilde açıklardı.	1	2	3	4	5	0
16. Öğretim elemanı derse dair önemli noktaları belirlerdi.	1	2	3	4	5	0
17. Öğretim elemanı kavramları açıklamak için güzel örnekler verirdi.	1	2	3	4	5	0
18. Öğretim elemanına ders dışında da ulaşılabilirdi.	1	2	3	4	5	0
19. Ödevlerin miktarı ve düzeyi öğrencilere uygundu.	1	2	3	4	5	0
20. Değerlendirme yöntemleri uygundu.	1	2	3	4	5	0
21. Ders, alanıma olan ilgimi artırdı.	1	2	3	4	5	0
22. Ders oldukça iyi organize edilmişti.	1	2	3	4	5	0
23. Ders materyali (okumalar vb.) faydalıydı.	1	2	3	4	5	0
24. Ders, alanıma ilişkin kavramları anlamamda faydalı oldu.	1	2	3	4	5	0
25. Ders benim için oldukça değerliydi.	1	2	3	4	5	0
26. **Öğretim Elemanının Genel Değerlendirmesi:** Daha önce ders aldığım öğretim elemanlarına kıyasla, bu kişi öğretim konusunda kesinlikle etkiliydi.	1	2	3	4	5	0

D. Turkish Summary / Türkçe Özet

Türkiye'de Bir Devlet Üniversitesinde Öğretim Elemanının Cinsiyeti ve Öğrenci Karakteristiklerine Göre Öğretim Etkililiğinin Algılanmasındaki Farklılıklar

Giriş

Araştırmanın Amacı ve Önemi

Bu araştırma, öğretim elemanlarının öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinde öğretim elemanının cinsiyetine ve öğrencinin cinsiyetinin, öğretim elemanından aldığı notun ve eğitim aldığı alanın incelendiği öğrenci karakteristiklerine göre farklılıklar olup olmadığını incelemek amacıyla yapılmıştır. Bu amaca yönelik olarak da çalışmada biri "Demografik Bilgiler" ve diğeri Young ve Shaw (1999) tarafından geliştirilen ve Emil (2013) tarafından Türkçe'ye uyarlanan "Öğretim Elemanı Değerlendirme Anketi" olmak üzere iki kısımdan oluşan bir veri toplama aracı kullanılmıştır.

Türkiye'de yükseköğretim kurumlarının bazılarında öğretim elemanlarının öğretim etkililiğinin değerlendirilmesinde öğrenci değerlendirme anketleri yaygın olarak kullanılmaya başlanmıştır. Bunda etkili olan faktörler şöyle sıralanabilir: Artan üniversite sayısıyla birlikte, kaliteyi güvence altına alma ve daha fazla öğrenciyi çekme kaygısının yeni bir yükseköğretim tüketimciliğine sebep olması (Harvey, 2011); üniversitelerin sınırlı bütçelerinin ve devamlı artan öğretim ücretlerinin öğretim elemanlarının etkililiğinin değerlendirilmesini gerekli kılması (Seldin, Miller and Seldin, 2010); öğrenci memnuniyeti, öğrenci sadakati ve öğretimin kalitesi arasındaki ilişki nedeniyle daha fazla öğrenci çekmek isteyen üniversitelerin öğretim kalitesini artırmaya yönelik çalışmaları (Lin vd., 2017); yükseköğretim kurumlarının öğrenci memnuniyetini artırmaya yönelik çalışmaları (Elez, 2017).

Orta Doğu Teknik Üniversitesi (ODTÜ) de bu uygulamaya sahip okullardan birisi olma niteliğini taşımaktadır. Dönem sonlarında uygulanan anketler kâğıt-kalem formatında ya da online olarak öğrencilere değerlendirmeleri için ulaştırılmakta ve sonuçlar yetkili personel tarafından üniversite, fakülte ve bölüm bazında takip edilmektedir. Temel Bilimler, Mühendislik, Mimarlık, Sosyal ve İdari Bilimler için doçentlik ve profesörlük atama ve yükseltme kriterleri arasında son 6 dönem ortalamasının 4.00'ten yüksek olması veya normalize edilmiş puanının fakültenin üst %80'lik dilimi içinde olması şartı aranmaktadır (ODTÜ, 2018). Değerlendirme sonuçları aynı zamanda öğretim etkililiklerine göre öğretim elemanlarının ödüllendirilmesinde ODTÜ Mustafa Parlar Vakfı tarafından da kullanılmaktadır. Mustafa Parlar Vakfı Ödül Yönetmeliğine göre ODTÜ Yılın Eğitimcisi Ödülü Rektörlük tarafından uygulanan bu Ders Değerlendirme Anketlerinin sonuçlarına göre verilir (ODTÜ Parlar Vakfı, 2018).

ODTÜ örneğinde görüldüğü gibi öğretim etkililiğinin artırılmasının yanı sıra ödüllendirme, atama-yükseltme gibi karar verme süreçlerinde de kullanılmalarına rağmen öğretim elemanlarının öğretim etkililiğinin öğrenciler tarafından değerlendirilmesi bazı soruları da beraberinde getirmektedir. Literatürde bu anketlerinin güvenilirlik ve geçerliğini sorgulayan çalışmalar, bu anketlerin çeşitli önyargılar taşıyıp taşımadığının araştırılmasını gerekli kılmaktadır.

İlk olarak, literatürde pek çok çalışma bu anketlerin cinsiyet önyargısı taşıyıp taşımadığını sorgulamaktadır. Basow, Phelan ve Capotosto (2006) 175 üniversite öğrencisinden en iyi ve en kötü olduklarını düşündüklerini öğretim elemanlarını betimlemelerini istemiş ve sonucunda en iyi öğretim elemanlarının %71'inin erkek olduğu ve en iyi ve en kötü kadın ve erkek öğretim elemanlarından beklenen özelliklerin cinsiyet stereotiplerine uyduğunu bulmuşlardır. Çalışmada en iyi kadın öğretim elemanları daha ulaşılabilir ve tutkulu bulunurken en iyi erkek öğretim elemanlarının daha bilgili, tutkulu ve yenilikçi olduğu öğrenciler tarafından rapor edilmiştir. Bu tür çalışmalar, anketlerdeki bazı maddelerin cinsiyet stereotipleri ile

uyuşabileceği ve çeşitli önyargılardan etkilenmiş sonuçlar üretebileceği konusuna dikkat çektikleri için önem taşımaktadırlar.

İkinci olarak, öğretim elemanının cinsiyeti ile birlikte öğrencinin cinsiyetinin de önyargılı sonuçlara neden olabileceği konusu bazı çalışmalar tarafından gündeme getirilmiştir. Das ve Das (2001) erkek işletme öğrencilerinin en iyi öğretim elemanı olarak bir erkek öğretim elemanını, kadın işletme öğrencilerinin ise bir kadın öğretim elemanını seçtiğini bulmuştur. Bu da öğrenci cinsiyetinin de öğrencilerin öğretim elemanlarını nasıl algıladıkları konusunda bir önyargı oluşturabileceği konusuna dikkat çeken çalışmalardan biridir.

Üçüncü olarak, literatürde öğretim elemanının dersinden alınan notun öğrenci değerlendirmelerini etkileyebileceğine yönelik çalışmalar mevcuttur. Lin (2015) öğrenciler ara sınav olup notlarını öğrendikten sonra uygulanan öğrencilerin öğretim elemanı değerlendirmeleri ile öğrencilerinin notları arasında bir ilişki bulmuştur. Ancak, öğrencilerin bu değerlendirmeleri doldurmadan önce not bilgisine sahip olmadıkları durumlarda ise aynı ilişki bulunamamıştır. Bu da öğrencilerin öğretim elemanı öğretim etkililiği değerlendirmelerini doldururken notlarını da göz önünde bulundurarak değerlendirme yapabilecekleri ve öğretim elemanının öğretim etkililiği konusunda değerlendirmelerde yanıltıcı sonuçlar olabileceğini gösteren çalışmalardandır.

Çalışma kapsamında son olarak öğrencilerin eğitim aldığı alanlara göre öğretim elemanı öğretim etkililiği değerlendirmelerinde farklılıklar olup olmadığı incelenecektir. Parpala, Lindblom-Ylänne ve Rytkönen (2011) üç farklı alanda öğrencilerin öğretim kalitesini nasıl kavramsallaştırdıklarını çalışmış ve farklı alanlardan öğrencilerin öğretimin kalitesini farklı şekillerde kavramsallaştırdığını bulmuş ve öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinde bu tarz farklılıkların olabileceğini göz önünde bulundurulması gerektiğini önermiştir. Benzer şekilde Spooren, Brockx ve Mortelmans (2013) de öğretim etkililiğinin kabul görmüş tek bir tanımının olmayışının öğretim etkililiğinin öğrenciler tarafından

değerlendirilmesinde kullanılan veri toplama araçlarının geçerliği konusunda tartışmalara yol açtığını dile getirmişlerdir.

Tüm bunlar göze alındığında mevcut çalışma, fakülte performans değerlendirme aracı olarak öğrencilerin öğretim etkililiğini değerlendirmesinin bazı yanıltıcı sonuçlar ortaya çıkarabileceğini (Anderson & Miller, 1997) göstermesi açısından önemlidir. Bu ve benzer çalışmalar değerlendirmelerin geçerliğini düşüren gözle görünmeyen önyargıların ortaya çıkarılması açısından önemlidir.

Ayrıca bu değerlendirme sonuçlarının sadece öğretim etkililiğinin artırılması değil, atama yükseltme ve ödül süreçlerinde de kullanılması mevcut çalışmanın yapılmasını önemli kılmaktadır. Sonuçların önyargılar taşıması durumunda uygulama süreçlerinin yeniden gözden geçirilmesi, akademik yöneticilerin de bu tarz çalışmaların hedef kitlesi olması (Kember & Ginns, 2012) çalışmayı önemli kılmaktadır.

Son olarak yükseköğretimde öğretim etkililiğinin daha iyi anlaşılabilmesi için de bu çalışma önemlidir. Neumann (2001) yükseköğretimde öğretim etkililiğinin anlaşılmasında disiplinler arası farklılıklar için hala açıklamalara ihtiyacımız olduğunu vurgulamıştır. Öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinin sonuçları göz önünde bulundurulurken de disiplinler arası farklı öğretim ve öğrenim ihtiyaçlarına dikkat çekmesi açısından da çalışma önem taşımaktadır.

Araştırma sorusu

Öğretim elemanının cinsiyeti ve öğrenci karakteristiklerine göre öğretim etkililiğinin öğrenciler tarafından değerlendirilmesi farklılıklar göstermekte midir?

Araştırma alt-soruları:

1. Öğretim elemanının cinsiyeti ve öğrencinin cinsiyetine göre öğretim etkililiğinin öğrenciler tarafından değerlendirilmesi farklılıklar göstermekte midir?
2. Öğrencinin notuna göre öğretim etkililiğinin öğrenciler tarafından değerlendirilmesi farklılıklar göstermekte midir?
3. Öğrencinin çalışma alanına göre öğretim etkililiğinin öğrenciler tarafından değerlendirilmesi farklılıklar göstermekte midir?

Literatür Taraması

Öğretim etkililiğinin öğrenciler tarafından değerlendirilmesi bazı yükseköğretim kurumlarında öğretim elemanı performans değerlendirmelerinde yaygın olarak kullanıldığı için yüksek öğretimde performans değerlendirmenin tanımı, amaçları ve uygulamalarının tartışılması önemlidir. Arnãutu ve Panc (2015) performans değerlendirmesini, sürekli ve standardize biçimde çalışanların performanslarının değerlendirilmesini içeren kurumsal bir süreç olarak tanımlar. Öğretim elemanı performans değerlendirmesi biçimlendirici ve özetleyici amaçlarla akademik yöneticiler ve öğretim elemanları tarafından kullanılabilir (Channing, 2016). Kalaycı (2009) birçok üniversitede öğretim elemanı performans değerlendirmesinin öğrenci değerlendirmeleri ile sınırlı olmasını ve sadece kalite güvence sistemlerinin rutin bir parçası olarak yürütülmesini eleştirirken uluslararası standartları yakalamak isteyen üniversitelerin daha iyi işleyen öğretim elemanı performans değerlendirme sistemlerine ihtiyacı olduğunu belirtir.

Literatür taramasının ilk kısmı öğrenciler tarafından öğretim elemanlarının öğretim etkililiğinin değerlendirilmesi üzerine tartışmalara odaklanmaktadır. Öğrenci değerlendirmeleri, öğretimin ve verilen derslerin değerlendirilmesi şeklinde kullanılan en yaygın öğretim elemanı değerlendirme şeklidir (Kember & Ginns, 2012). Kember ve Ginns, değerlendirmeyi karar verme süreçlerini doğrulamak için

veri toplama, inceleme ve yorumlama olarak tanımlar. Öğrencilerin öğretim etkililiğini değerlendirmesi de üniversitelerde hem biçimlendirici hem de özetleyici amaçlarla kullanılmaktadır (Puteh & Habil, 2011; Chalmers, 2011). Öğrencilerin öğretimi gözlemlemek için en yakın gözlemci olmaları ve öğretim eyleminin hedefi olmaları sebebiyle öğrenci değerlendirmeleri öğretim elemanı değerlendirme süreçlerinde değerli bir yere sahiptir (Pallet, 2006). Ancak değerli veriler sağladığı düşünülmesine rağmen öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinin öğrenme ve öğretim kalitesini artırdığı henüz tam olarak kanıtlanmamış (Kember & Ginns, 2012) ve öğrenme eyleminden ziyade tüketici bir bakış açısıyla öğretime odaklanması da eleştirilmiştir (Harvey, 2011). Benzer şekilde öğrencilerin öğretim etkililiğini değerlendirmesinde eleştirilen diğer bir konu da öğretim etkililiğinin tek ve her yerde geçerli bir tanımı olmamasıdır (Ryans, 1949; Arreola, 2007). Borich (2007) öğretim etkililiğinin farklı dönemlerde farklı tanımlara sahip olduğunun da altını çizerken güncel tanımlarda sadece öğretmen karakteristiklerinin değil öğretmen-öğrenci ilişkisinin de tanımın bir parçası olduğunu ifade eder. Öğretim etkililiğinin tanımının ortam şartlarına ve zamana göre değişmesine rağmen çeşitli çalışmalar araştırmacıların öğretim etkililiğinin tanımını yapmasına imkân sağlamıştır. Örneğin, Seldin (2006) öğretim etkililiğini değerlendiren araştırmaların tutarlı sonuçlar bulduğunu, etkililiği belirleyen unsurlar arasında derse hazırlıklı gelmek, öğretilen konu hakkında geniş bilgi sahibi olmak, öğrenci motivasyonunu artırmak, öğrencilere eşit ve adil davranmak, öğrenmenin detaylarına hâkim olmak, hem öğretmede hem de öğretilen konu hakkında istekli olmak gibi özellikler olduğunu belirtir. Seldin'e göre sorular uygun ve doğru şekilde sorulduğu takdirde, öğrenci değerlendirmeleri öğretimin etkililiği hakkında doğru ve güvenilir sonuçlar vermektedir. Ancak öğretim etkililiğinin kabul görmüş tek bir tanımının olmaması, farklı zamanların ve ortamların farklı öğretim ihtiyaçları olması bu öğretim etkililiği değerlendirme araçlarının geçerliği konusunda tartışmalara da yol açmaktadır (Spooren, Brockx, & Mortelmans, 2013).

Öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinde öğretim elemanının cinsiyeti, öğrencinin cinsiyeti, öğrencinin dersten aldığı not ve öğrencinin çalışma

alanı gibi faktörlerin değerlendirme sonuçlarına önyargı taşıyıp taşımadıkları ve değerlendirmelerin geçerliğini tehdit edip etmedikleri literatürde çeşitli çalışmalar tarafından araştırılmıştır. Literatürdeki bazı çalışmalar öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinde kadın öğretim elemanları aleyhine bir cinsiyet önyargısı olduğunu bulurken (Wagner, Rieger, & Voorvelt, 2016), değerlendirmelerin cinsiyet önyargısı taşımadığını söyleyen çalışmalar da vardır (Punyanunt-Carter & Carter, 2015). Öğrencinin dersten aldığı not ile öğretim etkililiğine verdikleri puanlar arasındaki ilişki de araştırılmış ve Cho, Baek ve Cho (2015) öğrencinin dersten beklediği not ile dersten aldığı not arasındaki farkın öğretim elemanı değerlendirmelerini olumlu etkilediğini bulmuşlardır. Araştırmacılara göre bu farkın öğrenci değerlendirmelerine yansıması öğrencinin bir şekilde öğretim elemanını aldığı daha yüksek not için ödüllendirme şeklidir. Cashin (1990) da farklı çalışma alanlarından öğrencilerin öğretim elemanlarını farklı şekilde değerlendirdiğini bulmuştur.

Literatür taramasının ikinci kısmı ise performans değerlendirmelerinde cinsiyet tartışmalarına odaklanmaktadır. Öğretim etkililiği ve öğretim etkililiğinin öğrenciler tarafından değerlendirilmesi önceki kısımda tartışıldığı için, bu değerlendirmelerdeki cinsiyet faktörüne daha detaylı odaklanmak gereklidir. Bu bölümdeki temel argümanları şunlar oluşturmaktadır: İlk olarak, bir iş yeri olan üniversiteler uygulamaları da cinsiyetçi olan cinsiyetli kurumlardır (Bird, 2011); ve ikinci olarak, üniversitelerdeki kişilerden toplumsal cinsiyet stereotiplerine uygun davranmaları beklenir (Lester, 2008) ve son olarak, performans değerlendirme süreçleri cinsiyetçi beklentilerden arınmış değildir (Acker, 1990). Performans değerlendirme sistemlerinde açıkça bir cinsiyet önyargısı gözlenmese bile Burton (1991) yaptığı çalışmada göstermiştir ki performans değerlendirme sürecinde verinin toplama şekli, soru maddelerinin ifade edilme şekli ya da veri toplama araçlarının dizayn edilme şekli cinsiyetçi olabilir. Burton'a göre bunlar verilerin sonuçlarını etkileyebilir ve kişilerin toplumsal cinsiyet ile alakalı algılarından ya da kadın egemen işlerin özelliklerine olan görmezlik durumundan kaynaklanabilir.

Burton (1991)'un bahsettiği gibi öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinde kullanılan veri toplama araçları da cinsiyetçi ifadeler taşıyabilir ya da aracın dizayn edilme şekli cinsiyetçi olabilir. Bu nedenle toplumsal cinsiyet perspektifi ile bu değerlendirmelerin incelenmesi gereklidir. Örneğin, MacNell, Driscoll ve Hunt (2014) online bir derste öğrencilerin öğretim elemanlarını değerlendirmesinde cinsiyet önyargısı taşıyıp taşımadıklarını araştırmışlardır. Kadın olarak sunulan öğretim elemanları, gerçek cinsiyetlerinden bağımsız olarak daha düşük puanlar almış ve çalışmada erkek öğretim elemanı olmanın profesyonellik, uzmanlık ve etkililikte otomatik kabul görmeyle ilişkili olduğu bulunmuştur. Ancak literatürde aynı zamanda öğretim etkililiğinin değerlendirilmesinde toplumsal cinsiyet önyargılarının olmadığına dair de çalışmalar bulunmaktadır (Fernández & Mateo, 1997; Price vd., 2017).

Sonuç olarak, tüm bu çalışmalar ve tartışmalar dikkate alındığında öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinde öğretim elemanının cinsiyeti ve öğrenci karakteristiklerine göre farklılıklar olup olmadığının araştırılması bu uygulamaların güvenilirlik ve geçerlik tartışmaları için önemli olmaktadır.

Yöntem

Desen

Mevcut çalışmada nedensel karşılaştırma araştırması kullanılmıştır. Çalışmanın verileri Orta Doğu Teknik Üniversitesi'nde eğitim almakta olan lisans öğrencilerinden toplanmıştır.

Örneklem

Türkiye'de öğrenim görmekte olan üniversite öğrencileri çalışmanın popülasyonunu oluşturmaktadır. Elverişlilik ve zaman gibi faktörlerin göz önünde bulundurulması nedeniyle çalışma Ankara'da Orta Doğu Teknik Üniversitesi'nde yürütülmüştür.

Çalışmaya 27 farklı bölüm ve 6 farklı alandan 667 öğrenci katılmıştır. Bu öğrencilerin 256'sı 2. sınıf, 191'i 3. sınıf ve 220'si de 4. sınıf öğrencisidir.

Veri Toplama Araçları

Veri toplama aracı iki kısımdan oluşmuştur: Demografik Bilgiler ve Öğretim Elemanı Değerlendirme Anketi. Veri toplama aracının "Demografik Bilgiler" kısmından öğrencilerden kaçıncı sınıfta oldukları, bölümleri, değerlendirmek üzere seçtikleri öğretim elemanının cinsiyeti, öğrencinin kendi cinsiyeti ve seçtikleri öğretim elemanından almış oldukları not ve genel not ortalamalarının bilgisi istenmiştir. Öğretim etkililiğinin değerlendirilmesi amacıyla ise de Young ve Shaw'ın 1999'da geliştirdiği ve Emil (2013) tarafından Türkçe'ye uyarlaması yapılan Öğretim Elemanı Değerlendirme Anketi kullanılmıştır. Veri toplama aracının bu kısmında öğrencilerden o zamana kadar dersini tamamlamış oldukları öğretim elemanları içinde en etkili ya da en az etkili öğretim elemanını seçip bu seçilen öğretim elemanının öğretim etkililiğini anketteki 25 soruyu cevaplayarak değerlendirmeleri istenmiştir.

Veri Toplama Süreci

Veri toplama süreci başlamadan önce Orta Doğu Teknik Üniversitesi İnsan Araştırmaları Etik Kurulu'nun izni alınmıştır. Veri toplama süreci 2017-2018 akademik yılının Bahar döneminde bir ay sürmüştür. Veri öğrencilerin sınıfları ziyaret edilerek toplanmıştır. Ziyaretlerden önce dersi veren öğretim elemanları bilgilendirilmiş, izinleri alınmış ve sınıfların ziyaret edileceği zamanlar kararlaştırılmıştır. Daha sonra veri toplama aracı öğrencilere gönüllü katılım esasına göre dağıtılmış ve öğrenciler veri toplama süreci ve aracın etik kurul izni konularında bilgilendirilmişlerdir.

Uygulama sınıflarda yaklaşık 15 dakika sürmüştür. Öğrenciler çalışma ile ilgili verdikleri cevapların gizliliği, diledikleri zaman ayrılabilecekleri gibi konularda da

uygulamadan önce bilgilendirilmiş ve araştırmacı veri toplanan sınıflarda bizzat bulunmuştur.

Veri Analizi

Çalışmanın verileri betimsel istatistikler ve çıkarımsal istatistikler kullanılarak IBM SPSS 23 programı aracılığıyla analiz edilmiştir. Data analizi gruplar arasındaki farklılıkları test etmek amacıyla nedensel araştırma yöntemi kullanılarak yapılmıştır. Birinci araştırma sorusu için iki ayrı faktör, öğretim elemanının cinsiyeti ve öğrencinin cinsiyeti, olduğundan iki yönlü ANOVA testi uygulanmıştır.

İkinci ve üçüncü araştırma soruları ise tek faktörlü oldukları için tek yönlü ANOVA testi uygulanmak istenmiş ancak varyansın homojenliği varsayımı ihlal edildiği için Welch's F test ile analize devam edilmiştir.

Araştırmanın Sınırlılıkları

Çalışmadaki veriler anket yoluyla toplanmış olup bulguların nedenleri hakkında bir nedensellik ilişkisi sunmamaktadır. Konunun daha derinlemesine anlaşılabilmesi için öğrencilerle ile öğretim etkililiğinin nasıl algılandığı konusunda kendi cinsiyetlerinin ve öğretim elemanının cinsiyetinin rollerine ilişkin mülakatlar yapılması faydalı olacaktır.

Çalışma Orta Doğu Teknik Üniversitesi'nde yapılmıştır. İngilizce'nin öğretim dili olması ve uluslararası kampüs ortamı nedeniyle öğrenciler popülasyonu temsil edebilecek düzeyde toplumsal cinsiyet önyargılarına sahip olmayabilirler.

Ek olarak öğretim elemanının etkililiğinin değerlendirilmesinde öğretim elemanının cinsiyeti, yaşı, unvanı, idari görevleri hepsi birlikte etkileşim içinde rol oynayabilirler. Ancak bu çalışma bu tür etkileşimlere odaklanmamış ve sadece farklılıkların betimsel ifade etmeye çalışmıştır.

Bulgular

Çalışmanın birinci araştırma sorusu "Öğretim elemanının cinsiyeti ve öğrencinin cinsiyetine göre öğretim etkililiğinin öğrenciler tarafından değerlendirilmesi farklılıklar göstermekte midir?" idi. Bu soru iki faktöre (bağımsız değişkene) sahiptir ve öğretim üyesinin cinsiyeti olan birinci faktörün kadın ve erkek olmak üzere iki seviyesi vardır. Aynı şekilde ikinci faktör öğrencinin cinsiyeti de 2 kadın ve erkek olmak üzere iki seviyelidir. Sorudaki bağımlı değişken ise öğrenci değerlendirmesindeki öğretim etkililiği skorudur. Yapılan iki yönlü ANOVA analizi öğretim elemanının cinsiyeti ve öğrencinin cinsiyeti arasında istatiksel olarak anlamlı bir etkileşim göstermemiştir. Bu etkileşim bulunmadığı için öğretim üyesinin cinsiyeti ve öğrencinin cinsiyeti için ayrı ayrı ana etkilere bakılmış, ancak her ikisi içinde istatistiksel olarak anlamlı bir ana etki bulunamamıştır.

Çalışmanın ikinci araştırma sorusu ise "Öğrencinin notuna göre öğretim etkililiğinin öğrenciler tarafından değerlendirilmesi farklılıklar göstermekte midir?" idi. Bu sorudaki bağımsız değişken öğrencinin notudur ve 12 seviyesi vardır: AA, BA, BB, CB, CC, DC, DD, FD, FF, W, S ve NA. 12 seviyesi olan tek bir bağımsız değişken olmasından dolayı tek yönlü ANOVA analizi planlanmış olsa da varyansın homojenliği varsayımı ihlal edildiği için Welch's F test ile analize devam edilmiş ve öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinde öğrencinin notuna göre istatistiksel olarak anlamlı farklılıklar bulunmuştur.

Çalışmanın üçüncü araştırma sorusu "Öğrencinin çalışma alanına göre öğretim etkililiğinin öğrenciler tarafından değerlendirilmesi farklılıklar göstermekte midir?" idi. Bu sorudaki bağımsız değişken öğrencinin çalışma alanıdır ve 6 seviyesi vardır: Mimarlık, Beşerî Bilimler, Mühendislik, Temel Bilimler, İktisadi ve İdari Bilimler ve Eğitim Bilimleri. Bu soruda da 6 seviyesi olan tek bağımsız değişken bulunduğu için tek yönlü ANOVA analizi planlanmış ancak varyansın homojenliği varsayımı ihlal edildiği için Welch's F test ile analize devam edilmiştir. Sonuçlar öğrencinin

çalışma alanına göre öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinde istatistiksel olarak anlamlı farklılıklar göstermiştir.

Çalışmanın ana araştırma sorusu "Öğretim elemanının cinsiyeti ve öğrenci karakteristiklerine göre öğretim etkililiğinin öğrenciler tarafından değerlendirilmesi farklılıklar göstermekte midir?" idi. Yukarıdaki analiz sonuçlarına göre çalışma, öğretim etkililiğinin öğrenciler tarafından değerlendirilmesinde öğretim elemanının cinsiyetine ve öğrencinin cinsiyetine göre istatistiksel olarak anlamlı farklılıklar göstermemekle birlikte öğrenci karakteristikleri olan ele alınan öğrencinin notu ve çalışma alanına göre anlamlı farklılıklar göstermiştir.

Sonuç ve Öneriler

Çalışmanın birinci bulgusu olan öğretim elemanının cinsiyeti ve öğrencinin cinsiyetine göre öğrencilerin öğretim etkililiğini değerlendirmesinde farklılıklar olmaması literatürdeki bazı çalışmalarla (Fernández & Mateo, 1997; Wheeless & Potorti, 1989) ile benzer sonuçlar gösterirken bazı çalışmalarla da (Basow, Phelan & Capotosto, 2006; MacNell, Driscoll & Hunt, 2014) farklı sonuçlar gösterilmiştir.

Orta Doğu Teknik Üniversitesi'nin çok kültürlü yapısı, İngilizce'nin eğitim dili olması, uluslararası kültürel değişimler için çok fazla fırsatlara sahip olması öğrencilerin sahip olduğu toplumsal cinsiyet stereotiplerini olumlu şekilde etkilemiş olabilir. Üniversitedeki bu çok kültürlü yapı ve çoğulcu ortam çalışmanın bulguları üzerinde etkili olmuş olabilir. İkinci olarak da Türkiye'de öğretmenliğin ideal bir kadın mesleği olarak görülmesi, Erginer ve Saklan (2016)'nın tanımladığı öğretmenlik mesleğinin kadınlaşması olarak literatürde bahsedilen eğitimci ve öğretmen kadın sayısındaki artış ve fazlalık da çalışmanın bulgularını etkilemiş olabilir. Çalışmanın ikinci önemli bulgusu ise not grupları arasındaki arasında bulunan istatistiksel olarak anlamlı farklılıklardır. Bu bulgular da öğrencinin not bilgisinin kesinliğine göre öğrenci değerlendirmelerinin sonuçlarını yönlendirdiğini söyleyen Cho ve Cho (2017) ile uyum içindedir. Yine çalışmayla uyumlu sonuçlara

ulaşan Eiszler (2002) de öğretim elemanlarının A ve A- notlarını daha fazla verdikleri dönemlerde daha yüksek öğrenci değerlendirmeleri aldığını bulmuştur. Çalışmanın üçüncü bulgusu ise Beşeri Bilimler öğrencilerinin değerlendirmelerinin Mimarlık ve Temel Bilimler öğrencilerinin öğretim elemanı etkililiği değerlendirmelerinden anlamlı olarak farklı çıkmasıdır. Bu da öğretim etkililiğinin geçerli herkes tarafından kabul görmüş tek bir tanımının olmamasından (Ryans, 1949; Arreola, 2007) kaynaklanabilir. Healey (2000) de öğretim kalitesinin öğretim ortamından bağımsız olmadığını söylemiş ve Parpala, Lindblom Ylänne ve Rytkönen (2011) de farklı alanlardaki öğrencilerin öğretim etkililiğini farklı şekillerde kavramsallaştırdığını bulmuştur.

Bu sonuçlardan yola çıkılarak üniversite yöneticileri, öğretim elemanları ve öğrenciler için aşağıdaki önerilerde bulunulabilir:

Çalışmada bulunan not ve çalışma alanına ilişkin sonuçlar nedeniyle üniversite yöneticileri özetleyici amaçlarla bu veri toplama araçlarını kullanırken dikkatli olmalı ve veri toplama araçlarını bunlardan etkilenmeyecek şekilde revize etmelidirler.

Akademik birimler kendi alanları için öğretim etkililiğinin ne demek olduğunu tartışmalıdırlar ve öğretim elemanları ile üniversite yöneticileri birlikte o kurum için öğretim etkililiğinin ne demek olduğunu yansıtan maddeleri öğrenci değerlendirmeleri için kullanılacak veri toplama araçlarına ekleyebilirler. Öğrenciler de bu veri toplama araçlarını doldurmanın duygusal yanı konusunda bilinçlendirilmeli ve bunların asıl amacının öğretim kalitesini artırmak olduğunu onlara söylenmelidir.

Öğrencinin not bilgisinin kesinliğinin öğretim etkililiğinin öğrenciler tarafından değerlendirmesi sonuçları etkileyebileceğinden üniversite yöneticileri bu veri toplama araçlarını bir kez dönemin ortasında bir kez de dönemin sonunda olacak şekilde uygulayabilirler.

Öğrenciler tarafından yapılan öğretim etkililiği değerlendirmelerindeki yüksek sonuçlar her zaman kaliteli bir öğretimin göstergesi olmayacağından sadece bunlara dayanarak yapılan öğretim elemanı ödüllendirme uygulamaları terk edilmeli ve öğrenci değerlendirmelerinin yanı sıra diğer değerlendirme metotları da sürece dahil edilmelidir.

Daha hafif iş yükü ve şişirilmiş notlar gibi etkenlerin etkisini azaltmak için öğrenci değerlendirmelerine dersin iş yükü ve öğrenci çabasını da değerlendirmeye yönelik maddeler eklenmelidir.

Öğretim ortamı ve dersin niteliği öğretim etkililiği değerlendirmelerini etkileyebileceğinden üniversite yöneticileri farklı nitelikteki derslerin ve öğretim elemanlarının etkililiklerini özellikle özetleyici amaçlarla kullanmak için karşılaştırmamalıdırlar.

Öğretim elemanları da öğrenci değerlendirmelerinin sonuçlarının öğrencinin notu ve çalışma alanına göre değişiklik gösterebileceğini bilmeli ve öğretim kalitesini artırmak için tek değerlendirme aracı olarak bu veri toplama araçlarını görmemelidir. Öğrencilerden sürekli yazılı ve sözlü geri dönütler alarak da öğretim kalitesini artırma yoluna gidebilirler.

Öğretim etkililiği değerlendirme sonuçlarını öğrencinin çalışma alanına göre de farklılıklar gösterebileceğinden öğretim elemanları öğretim kalitesini artırmak için öğretim ortamının şartları ve öğrenci ihtiyaçlarına göre de öğretimi şekillendirmeye istekli olmalıdırlar.

Öğretim elemanları kendi öğretim felsefeleri üzerine çalışmalı ve bununla ilgili bir farkındalık geliştirmelidirler. Zaman zaman sadece öğrencilerden değil öğretim ve ders materyalleri ile ilgili kendi meslektaşlarının da fikirlerine başvurabilirler.

Öğretim elemanları öğretim kalitesini artırmak için öğrenci değerlendirmelerinin önemini öğrencilerle paylaşmalı ve bu amaca yönelik veri toplama araçlarına öğrencilerin dürüst cevaplar vermesinin önemi konusunda öğrencileri bilgilendirmelidirler.

Literatür taraması öğrencilerin öğretim etkililiğini değerlendirmesinin önemini vurgularken öğrencilerin öğretimin bütün yönlerini değerlendirmede yeterli olamayacakları konusunun da altını çizmiştir. Bu nedenle öğretim elemanlarını öğretim kalitesini artırmaya yönelik değişiklikler planlarken bunu da göz önünde bulundurmalıdır.

E. Tez İzin Formu / Thesis Permission Form

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences	☐
Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences	X
Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics	☐
Enformatik Enstitüsü / Graduate School of Informatics	☐
Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences	☐

YAZARIN / AUTHOR

Soyadı / Surname : YILMAZ.....
Adı / Name : ECE.....
Bölümü / Department: EĞİTİM YÖNETİMİ VE PLANLAMASI / EDUCATIONAL ADMINISTRATION AND PLANNING .....

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English) : DIFFERENCES IN PERCEPTIONS OF EFFECTIVE TEACHING BY FACULTY MEMBER'S GENDER AND STUDENT CHARACTERISTICS IN A TURKISH PUBLIC UNIVERSITY

TEZİN TÜRÜ / DEGREE: Yüksek **Lisans** / Master ☒ **Doktora** / PhD ☐

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. ☒
2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two year. *** ☐
3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months. *** ☐

***** *Enstitü Yönetim Kurulu Kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir. A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.*

Yazarın imzası / Signature ..... **Tarih** / Date:.....