NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

THE IDENTIFICATION OF GENDER BIAS IN THE U.S. MILITARY

by

Luke T. Siwek
Brandon K. Wolf

March 2018

Thesis Advisor: Jeremy Arkes
Co-Advisor: Sae Young Ahn

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704–0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	March 2018	Master's thesis

4. TITLE AND SUBTITLE
THE IDENTIFICATION OF GENDER BIAS IN THE U.S. MILITARY

5. FUNDING NUMBERS

6. AUTHOR(S) Luke T. Siwek, Brandon K. Wolf

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB number NPS.2018.033-IR-EP7-A.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

Although females represent almost half of the U.S. civilian labor force, they account for less than 15 percent of the officers in the U.S. military. To account for this discrepancy, this thesis tests for gender bias within the U.S. military by analyzing unique datasets derived from Naval Postgraduate School. We first conduct a randomized control trial by means of a survey (n=234). One group responds to scenarios relating to one gender; the second group responds to the same scenarios but relating to the opposite gender. We then use statistical analysis and ordinary least squares models to compare responses between genders. Second, using NPS student evaluations of teaching (n=175,093), we conduct *t* tests, examine the correlation of evaluation questions on instructor effectiveness, and employ ordinary least squares models using student and course fixed effects, and instructor and course fixed effects while controlling for student, instructor, class and school characteristics to analyze how gender influences evaluations. Our results identify that students favor matched gender pairs, with the effect largest among male pairs. We found this effect to be of marginal economic significance. These findings may indicate the effectiveness of gender equality training, or may reflect the current social climate concerning gender bias.

14. SUBJECT TERMS	15. NUMBER OF PAGES
gender, bias, discrimination, military	119
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

NSN 7540–01-280-5500 Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

THE IDENTIFICATION OF GENDER BIAS IN THE U.S. MILITARY

Luke T. Siwek
Major, Australian Army
BEng(MechEng), Royal Melbourne Institute of Technology, 2002

Brandon K. Wolf
Lieutenant Commander, United States Navy
B.S., Miami University, 2003
EMBA, Naval Postgraduate School, 2017

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
March 2018

Approved by: Jeremy Arkes, Ph.D.
Thesis Advisor

Sae Young Ahn, Ph.D.
Co-Advisor

Yu-Chu Shen, Ph.D.
Academic Associate

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Although females represent almost half of the U.S. civilian labor force, they account for less than 15 percent of the officers in the U.S. military. To account for this discrepancy, this thesis tests for gender bias within the U.S. military by analyzing unique datasets derived from Naval Postgraduate School. We first conduct a randomized control trial by means of a survey (n=234). One group responds to scenarios relating to one gender; the second group responds to the same scenarios but relating to the opposite gender. We then use statistical analysis and ordinary least squares models to compare responses between genders. Second, using NPS student evaluations of teaching (n=175,093), we conduct *t* tests, examine the correlation of evaluation questions on instructor effectiveness, and employ ordinary least squares models using student and course fixed effects, and instructor and course fixed effects while controlling for student, instructor, class and school characteristics to analyze how gender influences evaluations. Our results identify that students favor matched gender pairs, with the effect largest among male pairs. We found this effect to be of marginal economic significance. These findings may indicate the effectiveness of gender equality training, or may reflect the current social climate concerning gender bias.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

BLS	Bureau of labor Statistics
CV	Curricula Vitae
DOD	Department of Defense
FI	Female Instructor
FITREP	Fitness Report
FS	Female Student
FSMI	Female Student Male Instructor
FSFI	Female Student Female Instructor
GSBPP	Graduate School of Business and Public Policy
GSEAS	Graduate School of Engineering and Applied Sciences
GSOIS	Graduate School of Operational and Information Sciences
IRB	Institutional Review Board
LPM	Linear Probability Model
MI	Male Instructor
MP	Must Promote
MS	Male Student
MSMI	Male Student Male Instructor
MSFI	Male Student Female Instructor
NAM	Navy Achievement Medal
NCM	Navy Commendation Medal
NPS	Naval Postgraduate School
OLS	Ordinary Least Squares
SD	Standard Deviation
SET	Student Evaluation of Teaching
SIGS	School of International Graduate Studies
SOF	Student Opinion Form
SSA	Social Security Administration
STEM	Science, Technology, Engineering, and Math
U.S.	United States
USN	United States Navy

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

We acknowledge and thank the NPS staff, and particularly our thesis advisors, for openly sharing their academic wisdom.

Luke further thanks his family for following him, and giving their full support in this endeavor. “To all the hard-working women out there that I know and love, this thesis is for you. To my mates back home—it’s time for a beer!”

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. BACKGROUND

Female participation rates within the U.S. military are below 20 percent, yet women make up almost half of the civilian labor force in the United States (Bureau of Labor Statistics [BLS], 2017). Furthermore, the female officer corps suffers a sharp drop off in the number of female officers above the rank of O4, to below 10 percent at senior levels (Defense Manpower Data Center, 2017). Clearly, any actions that sway either gender not to serve, or to leave service, can be devastating to recruitment, promotion, and retention efforts, as well as damaging to individual and organizational morale. Former U.S. Secretary of Defense Ash Carter once remarked that "to succeed in our mission of national defense, we cannot afford to cut ourselves off from half the country's talents and skills" (2015), preceded the opening of all military jobs to females in January 2016, and demonstrates the political importance of gender equality in the Department of Defense (DOD).

U.S. Military Manpower figures show a deficit in female representation when compared to their male counterparts, as shown in Table 1. Females make up too few of the DOD Officer Corps, contributing 15.3 percent of the total force; however, this contribution rapidly declines above the rank of O4. These figures are alarmingly different from what one would anticipate given the female labor force participation rate in the United States is around 75 percent, only 10 percent lower than the male participation rate, as shown in Figure 1. However, mirroring the civilian labor force, figures for senior female leadership in the U.S. military are very low; only 6.4 percent of Fortune 500 CEOs are female and 8.1 percent of O10s are female. We must ask ourselves why these gender differences exist.

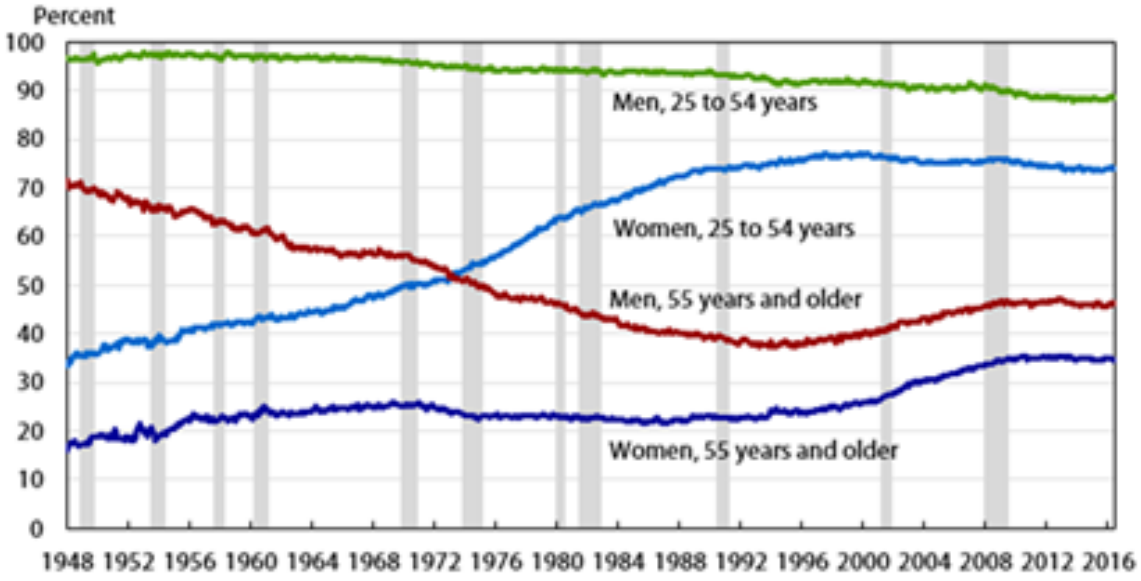
.

Table 1. Active duty female officers by rank and service.
Adapted from Defense Manpower Data Center (2017).

Aug 17 - Active Duty Females by Rank & Service							
	Army	Navy	Air Force	Marine Corps	Total	DoD Total	% Female
O10	0	1	2	0	3	37	**8.1%**
O9	3	4	5	0	12	150	**8.0%**
O8	6	5	8	1	20	309	**6.5%**
O7	8	10	7	0	25	417	**6.0%**
O6	478	352	485	16	1331	11132	**12.0%**
O5	1220	743	1527	78	3568	26842	**13.3%**
O4	2931	1657	2664	228	7480	42424	**17.6%**
O3	5699	4168	4863	461	15191	75609	**20.1%**
O2	2275	1461	1697	388	5821	28056	**20.7%**
O1	1715	1584	1695	294	5288	25586	**20.7%**
% Female Total	15.50%	15%	18.30%	5.80%			15.30%

Figures shown in red are below the average.

Figure 1. Labor force participation rates, 1948–2016.
Source: BLS (2017).



Note: Shaded areas represent recessions as determined by the National Bureau of Economic Research.
Source: U.S. Bureau of Labor Statistics, Current Population Survey.

Several economic factors may explain the decline in the representation of females among senior ranks. The age of achieving O4 may be when some females leave, or take a break, from the labor force to raise a family. Another

factor is that the vesting period (20 years) to receive the military's defined benefit, when coupled with the circumstance of raising a family, may also entice females to leave the military, knowing that they have some additional financial security. Or, as Table 1 could suggest, gender equality education is proving effective, reflected in the steady and higher representation of females at the more junior officer ranks. If, however, gender equality education is proving effective, we can expect to see the proportion of female officers remain steady across all ranks in the years to come. This thesis investigates another possibility that is increasingly drawing attention throughout society—the existence of gender bias.

While policy makers may be able to address some of the factors that possibly contribute to the comparatively low participation rates of females, particularly senior female officers, they are at risk of failing without first addressing gender bias. If gender bias exists it will affect decisions, countering the efforts of the best-intentioned policy changes. Therefore, it is in the interest of the DOD that gender bias research occurs, and if gender bias is found to exist, the DOD must develop strategies to address it.

Although gender bias can occur consciously and unconsciously, the methods we use to capture gender bias cannot definitively delineate whether the respondent is consciously or unconsciously displaying gender bias. We posit that if a military member overtly demonstrated gender bias, he or she would not survive within the organization. We believe the DOD promotes an environment of equality and, further, inculcates equality by training and educating individuals on diversity, as well as holding individuals accountable for biases. Nevertheless, research indicates that individuals can unknowingly demonstrate bias, hence the term unconscious. Decisions based on these biases, by leaders within organizations, can have far-reaching detrimental effects on an individual's career and work environment. If the DOD is not selecting the best people for the job due to unconscious biases, then ultimately operational effectiveness suffers.

What is the difference between unconscious and conscious bias? Unconscious bias, otherwise known as implicit social cognition, is a product of

how our brains process information, as well as how we are conditioned to respond to experiences (Greenwald & Banaji, 1995). This conditioning occurs throughout our lifetime and can be influenced through direct and indirect messaging. Individuals may not be aware of, or believe, that they possess such bias as it occurs at the subconscious or unconscious level of mental processing. However, even though implicit biases often occur "outside of a person's awareness, this does not preclude them from influencing behavior" (Snowden, 2005, pp. 4–5). Conversely, conscious bias, or explicit bias, refers to certain attitudes or beliefs an individual knowingly exhibits. Often blatantly exhibited, explicit bias is "strongly associated with racism" (Snowdon, 2005, p. 4) and is easier to distinguish by the affronted, and persons who come into contact with the offender. One important reason to distinguish between implicit and explicit biases is that it allows us to enlighten those who are unaware that biases can be formed, projected, and identified in more than one way. It is then the hope that this enlightenment will compel individuals to reflect on their attitudes and actions toward others.

If we discover gender bias, we will set the conditions for mitigation and further research toward closing the gender gap. Any efforts toward closing the gender gap can assist the U.S. military in coming closer to reflecting society. Moreover, fostering a gender equality environment can increase the talent available to the DOD, by no longer stalling opportunities for qualified individuals due to gender bias. A military freed of decision making influenced by bias naturally becomes a more effective and efficient organization with enhanced operational readiness.

B. PURPOSE OF STUDY

This study assesses whether gender bias exists within the U.S. military and, if so, to what extent. This is an important research topic as gender bias has the potential to impact recruitment, assignment, promotion, deployment, and

retention, as well as individual and organizational morale. With all military occupations opened up to females since 2016, this study is well overdue.

Members of the U.S. military are constantly assessed throughout their careers, by a multitude of assessors, and expect determinations free from bias. The effects of gender bias are not restricted to individual careers; when its effects are aggregated, it directly impacts the organization and operational readiness. If the DOD intends to enhance its force, it must eliminate any prejudicial influence on decisions and judgements.

Although service members expect their supervisors to make unbiased decisions and judgments, research tells us otherwise. As such, this thesis employs two complementary quantitative approaches to identify gender bias. First, we administer a randomized controlled trial: a small-scale survey, taken by a sample of Naval Postgraduate School (NPS) students. Additionally, we examine the NPS Student Opinion Forms (SOF), using numerous analytical methods, to further test for gender bias in the evaluation of instructors by the NPS student population. By identifying the existence of gender bias, and measuring its effect (if any), we take a large step toward allowing policymakers to design methods and subsequent programs to mitigate its adverse effect.

C. THESIS RESEARCH QUESTIONS

1. Primary Research Questions

Our primary research questions are:

a. Does gender bias exist within the U.S. military?

b. Does student gender influence the student's evaluation of an individual of the same or opposite sex?

c. If gender bias does exist, how prevalent is it?

2. Secondary Research Questions

Our secondary research questions are:

a. How do we measure gender bias?

b. What methods could be used to mitigate gender bias?

D. SCOPE AND LIMITATIONS OF STUDY

The scope and limitations of our study include:

1. Survey Analysis

The gender bias survey uses cross-sectional data drawn from a sample of current NPS resident Navy officers during the 2018 academic year. There are two main limitations to the survey research. As the survey was distributed only to active duty Navy personnel who are currently attending NPS, the survey was both restricted by the population and the correspondingly small sample size. These limitations may not allow for the results to be generalizable to all Navy active duty military officers and enlisted personnel. The student researchers and investigators considered this limitation when providing recommendations to policy makers who make decisions about Navy diversity training and education.

2. SOF Analysis

The SOF analysis utilizes a rich panel dataset that covers FY2007–FY2016 containing over 170,000 observations of a student in a given class. There are three main limitations to this dataset and our SOF research. First, we were unable to identify any SOF gender-related literature from a military perspective to inform our approach. This makes our thesis unique, filling a research gap. However, there is a plenitude of civilian sector research on student evaluations of teaching (SET), including many with a focus on gender bias. While these civilian studies may not be generalizable to the military, they do offer many insights for our analysis. Furthermore, as our population is unique, our findings

may not be generalizable to the civilian sector. Our findings may also not be generalizable across the wider military population.

Secondly, as the NPS student population is represented by junior officers, mainly at the rank of O3, it falls within the population of military officers where the proportion of females is stable across the rank range of O1 through O4. As discussed in the background of Chapter I, this population may already benefit from gender equality training, and if our research does not detect gender bias, then this limitation may just confirm the effectiveness of gender equality training.

The final limitation is that NPS SOFs evaluate the effectiveness of civilian instructors, not military officers. If NPS students hold professional civilians to different standards than they do military officers, our analysis of SOF data may not be generalizable to military members. The findings of this research only indicate NPS student evaluations of professional civilian staff, not necessarily military members. Nonetheless, we believe our findings are generalizable across the two populations; the findings may differ only in magnitude.

E. ORGANIZATION OF THESIS

This thesis comprises six chapters. Chapter I provides the reader with a cognitive framework to appreciate the wide and significant impacts of gender bias at the individual and organizational levels. It then focuses the reader on our purpose, testing for bias, before providing our research questions. It also describes the two distinct, yet complementary, approaches we take and their limitations.

Chapter II presents a review of selected prior studies that provide the framework for the development, implementation, and analysis of our gender bias survey. These studies focus on two types of field experiments, audit and correspondence studies, that inform us on the validity of utilizing said experiments in detecting bias and discrimination.

Chapter III offers the findings of contemporary literature related to student evaluations of teaching. It informs us of the numerous, conflicting research findings on the effects of gender bias on evaluations, along with the many methods employed to uncover gender bias. It also discusses the common research biases that plague SET research; the impacts of influential independent variables; and the many statistical and econometric methods used to uncover gender bias in evaluations.

Chapter IV discusses the data, methods, and results pertaining to our gender bias survey. First, we describe the survey question design, development, and distribution. We then detail the variables available to us and their summary statistics. This chapter also details the methods used to uncover gender bias as well as the results of said methods.

Chapter V discusses the data, methods, and results of the SOF analysis. First, it describes our rich panel dataset before detailing the variables available to us and their summary statistics. This chapter then details the methods used to uncover gender bias: *t* tests of mean responses by student and instructor pairs, SOF question correlations with question 12, and ordinary least squares (OLS) fixed effects models.

Chapter VI concludes our findings by answering our research questions, discussing the findings from our research methods (student survey and SOF data analysis), and provides recommendations for future research.

II. LITERATURE REVIEW: GENDER BIAS SURVEY

A. INTRODUCTION

Several research studies have looked at how conscious or unconscious biases affect hiring practices that can be characterized as discriminatory. Far too often, the conclusion of these studies centers on how attitudes, judgments, and stereotypes result in the rejection of an equally qualified applicant based on reasoning that has nothing to do with the applicant's qualifications or ability. In order to detect bias or discrimination, David Neumark (2012) explains that "earlier research on labor market discrimination focused on individual-level employment or earnings regressions, with discrimination estimated from the race, sex, or ethnic differential that remains unexplained after including many proxies for productivity" (p. 1130). Estimations, however, present some difficulties in that they "do not adequately capture group differences in productivity, in which case the "unexplained" differences cannot be interpreted as discrimination" (Neumark, 2012, p. 1130).

Field experiments such as audit and correspondence studies alleviate this weakness in the regression approach to capturing bias and discrimination. These study methodologies have provided "evidence consistent with discrimination, including discrimination against blacks, Hispanics, and women in the United States" (Neumark, 2012, pp. 1128–1129) and multiple other countries throughout the world. The field experiment approach used in audit and correspondence studies is instrumental to our gender bias survey development and analysis. Although our novel survey approach varies from these studies in that it is not a live actor performance, application, or resume, the key attributes of these methodologies are included.

A review of audit and correspondence studies gives a foundation for our survey and demonstrates the effectiveness of this approach in detecting bias and discrimination. With the knowledge gained from these approaches we gain a

broader understanding of the findings, implications, strengths, and weaknesses of current research, better equipping us for our purpose of testing for gender bias.

B. AUDIT STUDIES

One type of field experimental method utilized to ascertain and measure discrimination is an audit study. Audit studies have been used by urban and labor economists to study discrimination in the housing and mortgage market fields for many years (Yinger, 1995). Michael Fix and Raymond Struyk (1993) describe audit studies as experimental research that is conducted by sending out "two individuals (auditors or testers)" (p. 1) who have been "matched for all relevant personal characteristics," (p. 1) less one defining characteristic such as race, gender, or marital status, for which the researchers are attempting to detect an inequity. The auditors then apply for a job or purchase utilizing identical credentials except for the one defining characteristic. Finally, the "results they achieve and the treatment they receive in the transaction are closely observed, documented, and analyzed to determine if the outcomes reveal patterns of differential treatment on the basis of the trait studied" (Fix & Struyk, 1993, p. 1).

The advantage of an audit study as compared to other empirical methods is that audit studies can provide more direct evidence of discrimination (Neumark, Bank, & Van Nort, 1996). David Neumark and colleagues explain that other empirical methods sometimes infer gender discrimination in hiring practices by estimating gender differences in employment rates by controlling for the sex composition and other observed characteristics in the pool of applicants. These estimates may lead researchers to reach incorrect conclusions if differences between male and female applicants go unobserved (Neumark et al., 1996). However, "the audit methodology offers a potentially powerful means of overcoming" this issue as "unobservable differences between men and women are eliminated, at least in principle, by matching their characteristics" (Neumark et al., 1996, p. 917).

Nevertheless, audit studies do not come without their own limitations and biases. While the researchers go to great lengths to match auditor characteristics, the primary criticism of audit studies centers around the fact that although the auditors may be identical on paper, they do not appear identical to employers when the in-person interview takes place (Neumark, 2012). A further limitation of the audit method is that the "auditors know the purpose of the study," and therefore, they "may generate conscious or subconscious motives" to "generate data consistent or inconsistent with their beliefs about race or gender issues" (Bertrand & Duflo, 2016, p. 11).

For these reasons, and as NPS time constraints do not allow for such a study to be designed, practiced, and implemented in a span of two quarters, we chose not to utilize this method in our research. However, our survey format does utilize the deceptive nature of the audit study method to elicit responses that are true to the nature of the respondent. The importance of deception and its legitimacy is discussed further in Chapter IV, Section C, "Survey Question Development." In addition to the use of deception, the audit study method serves as the foundation for the development of our gender bias survey questions by establishing the need to make the fictitious officer characteristically identical and realistic.

C. CORRESPONDENCE STUDIES

Correspondence studies, sometimes referred to as resume audits, on the other hand address the criticism associated with audit studies by using "fictitious applicants on paper, or more recently the Internet, whose qualifications can be made identical across groups" (Neumark, 2012, p. 1129), while further reducing the complexity and limitations by not performing in-person interviews. This experimental research methodology builds upon audit studies and enhances our gender survey development, implementation, and analysis.

One major difference between an audit study and a correspondence study is that researchers "can only measure the interviewing stage" of the job or

purchase process (Lahey & Beasley, 2009, p. 2). However, Joanna Lahey and Ryan Beasley (2009) point out that as compared to audit studies, correspondence studies "allow the experimenter much more control over the experimental variables" while allowing the "experimenter to generate a large number of data points at a much smaller cost than does a traditional audit. Because resume audits allow for large sample sizes, large sample techniques can be used to analyze the data, providing more power and circumventing disagreements over which small sample technique is correct" (p. 2).

A key component to the correspondence study methodology is the development of identical resumes, applications, or curricula vitae, without key information such as gender and race that become the variables of interest. These studies attempt to elicit and measure bias from subjects without indicating the true nature of the study by distributing the fictitious applicant's resume, application, or other correspondence document via the intranet or mail. Our gender bias survey utilizes this methodology in a unique fashion by formulating identical survey questions, minus gender specific names and pronouns, which are based on realistic naval career scenarios and the respondent's perception of the appropriateness of the scenarios' conclusions. The following research serves as the basis for this approach.

1. Implicit Gender Bias in STEM Fields

Research conducted by Corinne Moss-Racusin, John Dovidio, Victoria Brescoll, Mark Graham, and Jo Handelsman (2012) analyzes and attempts to explain unconscious (implicit) gender bias via professor ratings of applications for a laboratory manager position. This research is important as females are underrepresented in science, technology, engineering, and math (STEM) related fields, with lifestyle choice being cited by some as the reasoning for this disparity. For this reason Moss-Racusin and colleagues (2012) utilized the correspondence method to explore if "given an equally qualified male and female student" would "science faculty members" demonstrate "preferential evaluation and treatment of

the male student to work in their laboratory" (p. 16474). They hypothesized that professors in the academic sciences would favor male students applying for the opening as a result of their view that males were more "competent" and "hireable" (Moss-Racusin et. al, 2012, p. 16475). Furthermore, the researchers posited that the professors would offer a higher "salary" and have a greater "willingness to mentor" male students who applied for the position (Moss-Racusin et al., 2012, p. 16475). The hiring of the student for the laboratory management position served as the dependent variable of interest, with secondary measures being "perceived student competence; salary offers, which reflect the extent to which a student is valued for these competitive positions; and the extent to which the student was viewed as deserving of faculty mentoring" (Moss-Racusin et al., 2012, p. 16475).

A sample of academic science professors (n=127) from across the United States, in the fields of biology, chemistry, and physics, were randomly assigned to review undergraduate students' applications for a laboratory management position. Unbeknownst to the professors, the applications that they were viewing were not from real students. This ruse was utilized to elicit implicit gender bias as the fictitious applications only varied by the name of the applicant. Recognizable and clearly gender-specific names, John and Jennifer, were utilized on the applications and all other information within the application was held constant; therefore, "any differences in the participants' responses" (Moss-Racusin et al., 2012, p 16478) were solely contributable to the gender of the applicant. Our current research borrows from this design by utilizing gender-specific names and pronouns, with much thought given to name selection, so as not to confuse the subject with androgynous names. Moss-Racusin and colleagues found that despite having identical qualifications, participants perceived applicants named John as significantly more competent, hirable, and offered them more mentoring than they did applicants named Jennifer. Additionally, the participants, who as scientists are trained to reject the subjective, were more likely to pay applicants named John a salary nearly $4,000 higher (13 percent) than they would

applicants named Jennifer. Interestingly, the researchers found that not only did male participants favor John over Jennifer, female scientists did as well. This is important as it demonstrates that the gender bias did not demonstrate in-group favoritism. That is, the negative bias toward females, or positive bias toward males, did not fall along demographic variables measured in the research (gender, scientific discipline, age, and tenure). Moss-Racusin et al. (2012) recognize that their results indicate a "subtle" or modest difference in ratings of male and female students; however, the impediments facing females within STEM-related fields can have "large, real-world disadvantages in the judgment and treatment of female science students" (p. 16477). Additionally, the researchers believe the bias that was measured was not generated from intentional, or conscious, bias. Rather, the bias demonstrated was most likely a result of a person's "repeated exposure to pervasive cultural stereotypes" (Moss-Racusin et al., 2012, p. 16474). Moss-Racusin et al. (2012) "reasoned that pervasive cultural messages regarding women's lack of competence in science could lead faculty members to hold gender-biased attitudes that might subtly affect their support for female (but not male) science students. These generalized, subtly biased attitudes toward women could impel faculty to judge equivalent students differently as a function of their gender" (p. 16475).

2. Implicit Gender Bias toward Job Applicants

Research conducted by Rhea Steinpreis, Katie Anders, and Dawn Ritzke (1999) examined whether applicants with identical curricula vitae (CV), apart from implied gender-specific names, would influence the academic institution's reviewers' decision to hire a job applicant or grant tenure to a candidate. To accomplish this, a group of male and female academic psychologists (n=238) were randomly selected to receive one of four versions of a CV (female job applicant, male job applicant, female tenure candidate, and male tenure candidate), and provide feedback about whether they would hire or grant tenure, respectively; and if applicable, to offer a starting salary amount (Steinpreis et al., 1999). The distribution method was selected in order to "limit the extent to which

the respondents would give politically correct answers" (Steinpreis et al., 1999, p. 512) as the researchers would only be answering questions about one gender. This method is similar to the one in our current study as each subject only received questions regarding either male or female career scenarios. Additionally, Steinpreis et al. (1999) developed the study's CV utilizing "standard information on the scientist's educational background, current institutional affiliation, teaching, research and service" (p. 514). Our study employs a similar approach in that the scenarios were developed to simulate real-life career experiences and circumstances in which naval officers may find themselves.

As with the Moss-Racusin study, Steinpreis et al. (1999) found that despite identical CVs, respondents were inclined to hire "applicants" with male names more so than "applicants" with female names (p. 520). Additionally, the researchers found that "both sexes reported that the male job applicant had done adequate teaching, research, and service experience"; whereas, "the female applicant" had not (Steinpreis et al., 1999, p. 509). Dissimilarly, the respondents who received the tenure candidate CV were "equally likely to tenure the male and female tenure candidates and there was no difference in their ratings of their teaching, research, and service" (Steinpreis et al., 1999, pp. 509–510), thus demonstrating an agreement as to the professional qualifications necessary to be considered for tenure. This difference between hire-ability and tenure-ability of a candidate brings into question when an applicant's professional experience, expertise, or record outweighs the gender bias discovered in the less experienced, yet identical, job applicant. As this research could not answer the question on what tips the "scales in terms of ensuring that a record is evaluated on its own merit rather than in light of the scientist's gender" (Steinpreis et al. 1999, p. 526), further research is recommended. Furthermore, the researchers acknowledge that educating staff on the existence of bias, as well as establishing of objective evaluation criteria, is needed to promote fair and equitable hiring practices.

3. Beyond Gender Bias: Bertrand and Mullainathan (2004)

To demonstrate the applicability of correspondence studies beyond gender discrimination, research by Marianne Bertrand and Sendhil Mullainathan (2004) titled, “Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on the labor market discrimination” is summarized here. This research demonstrates the need to perform comprehensive research within organizations, as individual (race, age, etc.) or group (religious affiliation, social-economic status, etc.) characteristics that lead to discriminatory practices extend beyond gender. As with gender, this discrimination also can significantly impact one’s ability to enter, advance, and remain with an organization.

Bertrand and Mullainathan’s research highlights the methods involved in conducting a correspondence study. To that end, Bertrand and Mullainathan point out that by using “conventional labor force and household surveys, it is difficult to study whether differential treatment occurs in the labor market” (2004, p. 993). Many institutions utilize these methods to collect demographic data, including the DOD via the Defense Manpower Data Center, which provides data to produce reports such as the annual Active Duty Member Demographics Report (Military OneSource, 2016). However, researchers who rely on survey data alone can usually only “measure differential treatment by comparing the labor market performance” of a particular characteristic “for which they observe similar sets of skills” (Bertrand & Mullainathan, 2004, p. 993). Conclusions based on these comparisons alone can be very misleading as these surveys do not account for observable differences from the employer’s perspective. A consequence of this disparity is that “any measured differences in outcomes could be attributed to these unobserved (to the researcher) factors” (Bertrand & Mullainathan, 2004, p. 993).

Although Bertrand and Mullainathan measure many other differences (e.g., response to resume quality and race) this literature review consciously focuses on race disparity in identical resumes as it most closely relates to our research and its analysis. As with the previously discussed gender

correspondence studies, Bertrand and Mullainathan's study relies upon the randomization of one key characteristic, that being race, while holding all other characteristics in the resume constant. The race on the fictitious resumes was expressed by assigning "very White-sounding names" to one half of the resumes and "very African-American-sounding names" to the other (Bertrand & Mullainathan, 2004, p. 992). The uniquely white or African-American names were determined through frequency data of birth certificates, a survey of respondents to attribute distinctiveness to the most frequent names, and then finally analyzed to determine the nine most likely names for each group (White-male, White-female, African-American-male, and African-American female). Our study utilizes a similar, albeit not as extensive, approach by analyzing the U.S. Social Security Administrations (SSA) database of the top 15 names of registered births since 2000 (see Chapter IV, Section B, "Survey Design"). Nearly 5,000 identical resumes, less the race characteristic, were sent to 1,300 employment ads in Chicago and Boston within an 11-month period. Responses were measured via a callback or email message for each fictitious resume. Results indicated that despite having identical resumes, there were statistically significant racial differences in employer response rates.

The results found that "applicants with White names need to send about 10 resumes to get one callback whereas applicants with African-American names need to send about 15 resumes," demonstrating a "50-percent gap in employer response to identical resumes" (Bertrand & Mullainathan, 2004, p. 992). Some may argue that callback rates in themselves are not proof enough that racial discrimination exists. Bertrand and Mullainathan state that

> [i]n a racially neutral review process, employers would rank order resumes based on their quality and call back all applicants that are above a certain threshold. Because names are randomized, the White and African-American resumes we send should rank similarly on average. So, irrespective of the skill and racial composition of the applicant pool, a race-blind selection rule would generate equal treatment of Whites and African-Americans. So our results must imply that employers use race as a factor when reviewing resumes,

> which matches the legal definition of discrimination. (Bertrand & Mullainathan, 2004, p. 1006)

D. CONCLUSION

This literature review of audit and correspondence studies serves as a point of reference for the development, implementation, and evaluation of the gender bias survey portion of our thesis. We revealed that the underlying methodology for these research approaches is the establishment of identical resumes, CVs, and/or personas, minus one specific individual characteristic (gender, race, etc.), in order to capture bias and discriminatory hiring or purchase practices. In addition to identical resumes, the literature provides a standard for the unorthodox use of deception. That is, the literature substantiates the practice of creating a persona in real time or on paper around a particular characteristic (gender) that can be manipulated through randomization and measured via statistical analysis by comparing responses by way of gender pairs.

This literature review also reveals that bias, although understated, is perceptible and measurable through the use of these methodologies. This knowledge provides solid footing for the development, implementation, and analysis of our survey data, and ultimately the capacity to uncover whether gender bias exists among the NPS naval officer student population.

III. LITERATURE REVIEW: STUDENT EVALUATIONS OF TEACHING

A. INTRODUCTION

Student evaluations of teaching (SET) began as early as the 1920s (Mau & Opengart, 2012) and have since become a widely accepted practice in higher education. Not only are SETs an accreditation requirement for higher education institutes across the world, institutions also use them to measure teacher effectiveness. SETs aid decisions about promotion, tenure, contract review, and the issuing of awards. NPS is no exception, employing the use of student evaluations under the title student opinion forms, or SOF.

This literature review first examines the breadth of SET research to gain an understanding of the range of topics covered and methods of analysis used. This examination informs us of the relationship between variables and SETs and the wide range of approaches we can employ to answer our research questions. We then specifically review research relating to gender bias and SETs, and learn that results are often disparate, largely due to research design. To enable us to more accurately capture gender bias in our research, we next explore the issues that complicate SET research and discuss common research biases. Similarly, we discuss the most influential variables, determined by research, and assess their effects on SET ratings. Consideration of these variables, when building our models, allows us to achieve more accurate coefficient estimates. Finally, we review SET research and theories on gender bias from social psychologists, which show how associating perceived stereotype characteristics to evaluation question themes helps uncover evaluation differences between males and females. We were unable to uncover any military SET gender-related studies during our literature review; however, we build on the numerous SET studies in the civilian sector to answer our research questions.

B. SET RESEARCH

Given the influence of these ratings on instructor careers, SET ratings have been researched widely by economists and social psychologists. Research topics are also extensive, analyzing the numerous relationships that variables and their interactions have on SET ratings. Research investigates: the validity and reliability of SET scores (Gilmore & Greenwald, 1999), demographic influences (Basow, Codos, & Martin, 2013; Bennett, 1982; Boring, 2017; Centra & Gaubatz, 2000; Maricic, Djokovic, & Jeremic, 2016; Reid, 2010), and educational characteristics, such as class size, class type, and discipline (Bedard & Kuhn, 2008; Centra & Gaubatz, 2000; Krueger, 2002; Marsh, Bornmann, Mutz, Daniel, & O'Mara, 2009; Potvin & Hazari 2016). Research into what influences SETs has even considered the time of day and weather conditions (Braga, Paccagnella, & Pellizzari, 2014), as well as the number of rows in a classroom (Safer, Farmer, Segalla, & Elhoubi, 2005).

The analysis methods used to research SETs also come in many forms. Research from a social psychology perspective favors statistical analysis of data, such as the means and standard deviations of responses, contingency tables, chi-squared tests, statistical hypothesis testing, and the range of univariate and multivariate analysis of variance tools. Economists, by contrast, favor elaborate multivariate econometric models, including ordinary least squares (OLS), logistic, ordered and ordered logistic proportionate odds models, incorporating student or instructor fixed effects, and in few instances both.

A review of SET literature gives an appreciation of the quantity of research and diverse methods of analysis conducted. With this knowledge, we gain a broader understanding of the findings, implications, strengths, and weaknesses of current research that can better equip us for our purpose of testing for gender bias.

C. SET AND GENDER DIFFERENCES

As we turn to our purpose of testing for gender bias within U.S. military, and review directly relevant literature, we uncover disparate results: some researchers find that there is evidence of gender bias, but others find no evidence of gender bias. We also find numerous researches asserting that gender bias studies produce conflicting results (Arbuckle & Williams, 2003; Basow, 2000; Centra & Gaubatz, 2000; MacNell, Driscoll, & Hunt, 2015).

MacNell et al. (2015) further discuss the conflicting research findings and propose reasons that may explain the disparity in results of gender differences, stating that "inconclusive results may lie in the research design of these previous studies" (p. 295). When we consider the multidimensional nature of SET analysis, it becomes apparent that without due consideration of potentially biasing factors, results are likely misinformed and misleading. MacNell et al. (2015) highlight the need for further scrutiny when assessing the validity of findings. We must ask ourselves: how was the experiment designed? What specifications were used? What interactions were created? How were known biases handled? And, under what context were findings discovered?

Narrowing literature to studies directly related to gender bias and SETs uncovers conflicting results, reinforcing the important link between experimental design and reliable results. Having briefly reviewed both general SET research and those studies directly related to gender bias, we must now consider those areas found to affect the reliability of previous findings—research biases and choice of independent variables.

D. COMMON FORMS OF RESEARCH BIAS

Natural experiments assessing student satisfaction of teaching are multidimensional and subject to numerous research biases. Unless we account for these biases, results will be misleading. Our appreciation of current SET research allows us to uncover the most common research biases affecting SET research: inadequate bias to judge, self-selection bias, and nonresponse bias.

Understanding the detrimental effects of these biases allows us to strategize mediating methods—aiding our overall research design.

A widely accepted finding is that SETs capture student perceptions of teacher effectiveness, not actual effectiveness (Alauddin & Kifle, 2014; Basow et al., 2013; Bursdal, 2008). Allauddin and Kifle (2014) argue that students "are not fully informed customers" (p. 5), and have an inadequate basis to judge; students primarily attend class to learn what instructors have to offer, not necessarily to assess teaching effectiveness. Boring (2017) takes this one step further and cites Statistical Discrimination Theory (Arrow, 1973; Phelps, 1972) to propose that when students are not fully informed to make an evaluation, they will fill information gaps with group averages based on what they do know (stereotypes). Solving this problem by increasing student knowledge of teacher characteristics is impractical. However, assuming some teacher characteristics are time-invariant (effectiveness, ability, motivation, etc.), the use of teacher-fixed effects in our modeling can solve the issues of omitted variables bias caused by these characteristics, as they drop out of the model.

Self-selection bias plagues many natural experiments. In the context of SETs, students who can choose their courses and instructors will likely produce more favorable evaluations. Potvin and Hazari (2016) studied gender bias in student evaluations of physics teachers and found students' subject association to have a large bias on evaluations. Specifically, they found "students with a strong physics identity show a larger gender bias in favor of male teachers than those with less of a physics identity" (Potvin & Hazari, 2016, p. 1). Similarly, Wright and Jenkins-Guarnieri (2012) citing Marsh and Dunkin (1997) in their meta-analysis of student evaluations of teaching literature "found that prior subject interest was the variable most strongly correlated with SETs" (p. 685). A method to reduce the effects of self-selection bias has been to limit sample populations to those students attending their first year of study (Boring, 2017; Carrell, Page, & West, 2010; Hoffman & Oreopoulos, 2009). In these instances, students attend classes in common general subjects and are randomly assigned

to instructors, which affords greater credibility to their findings. Students at NPS are randomly assigned to instructors, which limits the impact of self-selection bias in our sample.

Another bias plaguing SETs is nonresponse bias. Nonresponse bias occurs when there are differences in ratings between those who complete the evaluation and those who do not. Nonresponse bias, therefore, leads to data not representative of the population. Reisenwitz (2016) found that "there are significant differences between those who complete online student evaluations and those who do not" (p. 7). Unfortunately, many higher educational institute policies surrounding the administration of SETs deem them voluntary in nature and, therefore, subject to nonresponse bias. As NPS requires students to complete their SOF to obtain their grade, our data is free from nonresponse bias.

Biases challenge the accuracy of natural experiments. Acknowledging and understanding an inadequate basis to judge, self-selection bias, and nonresponse bias emphasizes the importance of correcting for, or eliminating, their influence. However, limiting our approach to deal solely with research biases is insufficient; the choice of variables used also has a large impact on the accuracy of results.

E. VARIABLES OF INTEREST FROM LITERATURE

The multidimensional aspects that influence student evaluations are complex. Thus, choosing what variables to include, when developing an econometric specification, is critical to the validity, reliability, and accuracy of the results. Understanding the impacts of the inclusion and omission of variables on results requires its own analysis and will further improve the validity of our research. We now discuss those variables that research has determined as having a significant impact on SET results. They are academic discipline, class size, academic rank, and student grade.

Academic discipline is an important aspect that requires inclusion in analysis. Otherwise, average effects across disciplines may dilute true coefficient

estimates. This was a strength of Potvin and Hazari's (2016) experimental design, as their study focused on physics; however, their findings are not generalizable. Reid (2010) reinforces the importance of inclusion of academic discipline. He finds that evaluations favor women less in traditionally masculine disciplines, and more favorably in traditional female disciplines. These findings synchronize with role congruity theory; that is, how well one's characteristics align to the perceived characteristics of one's role (Eagly & Karau, 2002). Although Reid (2010) articulates this concept well, his analysis uses data from ratemyprofessor.com to examine the effects of perceived race and gender of the instructor, which introduces self-selection bias. The diverse spread of academic disciplines at NPS, four schools overseeing 14 academic departments, provides us the opportunity to analyze gender effects between disciplines, and thus, avoid coefficient estimate dilution generated from a whole-of-school analysis approach.

Class size also influences SETs. Bedard and Kuhn (2008) found "a large, highly significant, and nonlinear negative impact of class size on student evaluations of instructor effectiveness that is highly robust to the inclusion of course and instructor fixed effects" (p. 253). Opposite findings were found by Liu (2012) in her examination of student and instructor characteristics and their influence on SET ratings. However, the context of her research was distance learning, where one could predict class size to have minimal impact on student evaluations of teaching. While a possible impacting variable, NPS has class sizes with little variation across disciplines. The advantage to this small variation will see class size have a minor effect on our results, allowing us to capture the effects of our other variables more accurately.

Another variable of significance has been academic rank. Liu (2012) found that "compared to Instructors, Assistant Professors and Professors tend to receive lower ratings on multiple dimensions, while Associate Professors are rated as high as Instructors" (p. 479). It may be fair to assume that as her study was for online courses, many other demographic biases may be eliminated as they may not be disclosed to students, reinforcing the importance of academic

rank to SET rating. This finding may also explain why other studies find little impact of academic rank as its effect is washed out by other evaluator biases.

The influence of student grade on SET rating has been shown to suffer from the "leniency hypothesis." This hypothesis describes a case of reverse causality in which instructors allocate higher grades to receive higher ratings (Wright & Jenkins-Guarnieri, 2012, citing Greenwald & Gilmore, 1997; Grump, 2007). Alauddin & Kifle (2014) also support a positive association with grade and SET rating. Addison, Best, and Warrington (2006) investigated this relationship further and when controlling for grade they found students' assessments on how difficult they thought a course was, in comparison to their expectations, positively correlated with professor evaluation. Comm and Mathaisel (1998) discuss how this also allows instructors to take advantage of SETs by teaching for favorable evaluation, not necessarily for student learning—a common critique of SETs.

Reducing the effects of biases that have troubled previous studies and including considered key variables brings results closer to causal relationships. As noted by Arbuckle and Williams (2003) citing Wilson (1998), Basow (1995, 2000) and Unger (1979), "parsing and analyzing particular items in a multidimensional evaluation instrument can pinpoint potential biases in the evaluation instrument, or, more notably, the implicit biases of the student evaluators" (p. 508). Here, we find that undue consideration toward the inclusion of academic discipline, class size, academic rank, and student grade will likely skew estimates away from their true influence on SET rating. However, and similarly with solely addressing research biases, stopping at variable consideration is not enough. To get closer to causal effect, we must turn to what social psychologists have found to contribute to gender biases in SETs.

F. UNCOVERING GENDER BIAS: SOCIAL PSYCHOLOGICAL FACTORS

In addition to understanding research biases and influential variables in previous studies, it is important to study the social psychological aspects of gender bias. This understanding can help us design suitable econometric

specifications and validation methods. MacNell et al. (2015) citing Monroe, Ozyurt, Wrigley, and Alexander (2008) articulate the importance of such an approach stating, “an examination of gender bias in student ratings of teaching must be framed within the broader context of the pervasive devaluation of women, relative to men, that occurs in professional settings in the United States” (p. 293). We will discuss the significance of role congruity and gender congeniality, gender expectations, gender stereotypes, differing acceptance and confirmatory standards, and how they affect student evaluations. An understanding of these influences on student evaluations assists our own experimental design as they offer insights into how we may identify gender bias within our sample population.

While defining gender bias is simple, it is the question of why gender bias occurs that makes cleanly identifying it challenging. Eagly and Karau (2002) posited the strong influence of role congruity and its effects on female leaders. Garcia-Retamero and Lopez-Zafra (2006), in a similar topic of research discussed the impacts of role congruity[1] of a leadership position, in a male congenial environment.[2] Their study sampled both male and female evaluators from undergraduate and high school students, workers, and retirees. The audience assessed the likelihood of promotion for a candidate (randomized gender), in a leadership role, across three industries: male congenial, female congenial, and not specified. They found that differences in the assessments of leadership are impacted by the (in)congruity of the evaluator’s gender and perceptions of leadership stereotypes. Further, the gender congeniality of the working environment also influenced evaluations. Garcia-Retamero and Lopez-Zafra (2006) revealed that if the working environment was congruent with a specific gender, then results were most likely favorable for that gender. These results are alarming. They suggest an individual must fit both a role-driven

[1] Role congruity theory describes how individuals or groups are evaluated positively when their characteristics align to the perceived stereotypes of the evaluator.

[2] An environment that aligns to a gendered stereotype (i.e., policing-male, nursing-female).

gender stereotype and the majority gender representation of the environment to be perceived as successful in comparison to an equally able individual of the opposite gender. Such findings clearly reinforce the hypothesis of the glass ceiling.

Foschi (2000) states, “experimental research provides clear evidence of stricter standards for women than for men when both perform at the same level and performance evaluations are objective” (p. 39). This theme continues to resonate among other researchers (Arbuckle & Williams, 2003; Basow, Phelan, & Capotosto, 2006; Garcia-Retamero & Lopez-Zafra, 2006; Kierstead, D’Agostino, & Dill, 1988; Potvin & Hazari, 2016). These differing standards are well articulated by Maricic et al. (2016) citing Anderson and Smith (2005) and Vilian (1998) regarding female teachers, “Students seem to expect more nurturing behavior from them, but they, in turn, often judge that behavior to be less professorial. Contrarily, if women teachers fail to meet students’ expectations of women, they are characterized as too masculine” (p. 198). Understanding these elements helps to define how we may test for gender bias by looking more closely at the relationship between stereotype characteristics and SET evaluations. What is most alarming about gender stereotyping is how it more negatively affects a female when she deviates from the perceived stereotype in comparison to a male.

If we now consider those researched areas to show favor to a specific gender, we find those displaying more feminine-communal characteristics will improve the evaluation for a female, whereas those displaying more masculine-agentic characteristics will improve the evaluation for males (Eagly & Karau, 2002). Given this theoretical context, a female leader in a male-congenial environment (i.e., higher education and military), on average, is most likely to succeed if she displays characteristics in line with how a female leader is expected to behave, not how a male leader is expected to behave. Role congruity theory, therefore, adds another layer of complexity that suggests success as a leader is determined by how well one meets the stereotype

perception of that gender, not a gender-neutral leader. This raises the question of how military personnel perceive their leadership whether within the military or the NPS academic environment. Is it one of gender-neutral characteristics, given females self-select into a male-congenial environment and, therefore, are perhaps accepted and equally matched to the job with their male peers, or are military personnel equally susceptible to gender biases as are their non-military counterparts given the male-congenial nature of the environment? MacNell et al. (2015) echo students' differing standards based on perceived stereotypes, stating that students "expect their male and female professors to behave in different ways or to respectively exhibit certain 'masculine' and 'feminine' traits" (p. 294). They then define professionalism and objectivity as male characteristics, and warmth and accessibility as female characteristics. Clearly, these traits deserve consideration in our analysis.

Biernat, Fuegen, and Kobrynowicz (2010) further complicate the dynamics of gender bias, finding that there are also differences in minimum acceptance and confirmatory standards for the same assessment, when gender and other role incongruity characteristics are considered. As an example, Biernat et al. state, "minimum standards for competence in 'masculine' occupations are lower for women than men, but confirmatory standards … are higher for women than men" (p. 855). Their research consisted of two studies concerning gender differences: evaluations of the number of behaviors assigned to suspect incompetence and the total number of behaviors to confirm incompetence; and the evaluation of a resume (content identical) for a generic officer position, with half the participants receiving the resume for a masculine job title (Executive Chief of Staff), the other half female (Executive Secretary). Their results show that again gender bias is likely prevalent when stereotypes impair our ability to evaluate fairly.

If we now consider the purpose of this study, particularly in the military and academic context, we can conclude that females in leadership roles (incongruent), in a male-congenial workforce, are highly likely to have their

competence assessed under far greater scrutiny than the normative group; that is, their male peers (Eagly & Karau, 2002). Therefore, for a female to succeed, she must play to the strengths of what an individual stereotypically looks for in a female leader. This notion is well supported (Arbuckle & Williams, 2003; Basow et al. 2006; Boring, 2017; Foschi, 2000; Kierstead, D'Agostino, & Dill, 1988; MacNell et al., 2015; Maricic et al., 2016). Reid (2010), however, did not find any main effect of gender until it intersected with race. These findings again reinforce the importance of understanding selective perceptions and stereotyping and their impact on evaluations, again emphasizing the importance of analyzing possible differences in characteristic assessments by gender.

Role congruity and gender congeniality, gender expectations, gender stereotypes, and differing acceptance and confirmatory standards are all closely related and point to the importance of including some form of gender-related characteristic traits into our models for analysis. Boring (2017) convincingly addressed the detection of gender bias in her approach by complementing her SET analysis with an examination of how the gender evaluations differed across the dimensions of teaching; a proxy for gender stereotypes and characteristics. The logic of this approach is that some aspects of the teaching dimensions are more aligned to stereotypical female characteristics (interpersonal traits, such as warmth and accessibility) and others more aligned to male stereotype characteristics (effectiveness traits, such as professionalism and objectivity) (MacNell et al., 2015). Therefore, grouping evaluations to stereotype expectations proves a viable method to uncover (un)conscious gender biases. This is a convincing method given the evidence of the effects on evaluations spanning gender, stereotypes, role congruity, and gender-congenial environments.

G. CONCLUSION

This comprehensive literature review of contemporary studies has included the analysis of a range of general SET studies, as well as SET gender

bias specific studies. We uncovered that analyzing statistical differences between responses across the four student and instructor gender pairs may provide misleading results, which helps explain the disparity in research findings. We then discussed research biases likely to affect experimental design and the key variables that skew findings. To best uncover gender bias, we must consider the impacts of research biases and those variables researched to heavily influence results when designing our analytical methods.

This literature review has also examined what social psychologists attribute as the root causes of gender bias in SETs—stereotyping of gender and role characteristics, the effects of gender-congenial environments, and differences in confirmatory and acceptance standards. This knowledge can assist our approach in detecting gender bias within our sample. The multidimensional aspects of gender bias also show how analyzing for gender bias by aggregating SET ratings is flawed, and that the multidimensional nature of SET ratings requires an appreciation of perceived gender teaching characteristics, if we are to accurately detect gender bias. It is these insights that steer our approach to uncovering gender bias in the evaluation of SOFs.

IV. DATA, METHODS AND RESULTS: GENDER BIAS SURVEY

A. INTRODUCTION

This chapter builds upon the findings of our gender bias survey literature review and discusses question design, survey development and distribution, as well as our methods of analysis. We first provide our Institutional Review Board (IRB) approval for human subject research, the rationale behind our research design, and the elements used in the development of our questions. Having established the basis for our questions we then discuss the format of the survey to include the deceptive nature of the title and first consent. We next introduce our dataset before discussing our variables, their description, and summary statistics. Having provided context for our survey design, development, and distribution, we turn to answering our research questions by discussing the methods used to uncover the existence of gender bias in our population and the subsequent results. We then use OLS regression to test for gender bias among our population by looking for differences, by participant and survey question gender pairs, in obtaining one of the seven ordinal ratings on a Likert scale.

B. IRB APPROVAL

This research involved the use of human subjects; therefore, approval was obtained from the NPS IRB as well as from the NPS President in order to recruit and collect information on participants. This thesis received IRB approval on February 22, 2018. The approved review protocol number is NPS.2018.033-IR-EP7-A.

C. SURVEY DESIGN RATIONALE

The design of our gender bias survey seeks to identify the existence of gender bias by surveying a sample of the NPS student population. The survey design consists of identical questions developed of the same nature, with the only difference being the gender of the name associated with the question and

gender-specific pronouns. Such a design, coupled with random allocation to respondents, allows us to attribute differences to gender bias. As outlined by Bertrand and Mullainathan (2004), our survey design thwarts the weaknesses of audit studies by creating scenarios in which indistinguishable qualities and traits of fictitious naval officers, minus the gender variation, are realistic and believable. By relying on scenario-based questions versus actors to portray officers in person, we can be sure that we are only comparing across gender and not an unknown quality that may be portrayed in a live performance (see Appendix C for survey question content).

D. SURVEY QUESTION DEVELOPMENT

The first step in the design of our survey was to design two questions about which all U.S. Navy officers would have a general understanding. Two themes that junior and senior officers alike are very familiar with are fitness reports (FITREP) and personal awards. These themes also allow for the formulation of realistic scenarios from which the respondents (U.S. Navy officers) could formulate a timely response.

In addition to the theme, the respondents could not realize that the questions are attempting to ascertain gender bias, as this would undermine the findings. Therefore, the questions are posed in a manner so as not to 'tip off' the respondent as to the true nature of the survey. This technique may be construed by some as deceptive as the technique is "contrary to the ethical standards for research established by the federal government" (Pager, 2007, p. 126). However, the policies that govern "the protection of human subjects" also recognize "that certain types of research" (Pager, 2007, p. 126) cannot be appropriately conducted by obtaining informed consent, and therefore, failure to acquire informed consent is acceptable in certain cases. Specifically, regulations allow for the waiving of "informed consent provided (1) the research involves no more than minimal risk to human subjects; (2) the waiver or alteration will not adversely affect the rights and welfare of the subjects; (3) the research could not

practicably be carried out without the waiver or alteration; and (4) whenever appropriate, the subjects will be provided with additional information after participation" (Pager, 2007, p. 126).

In instituting said deceptiveness, by way of the NPS IRB approval, the aforementioned criteria for the "misleading" initial informed consent was met as the participants were fully informed of the survey's true intent after it had been closed via a follow-on email. The follow-on email served as the "true" informed consent and instructed the participants that they could withdraw from the study and their data would be excluded (see Appendix D for follow-on email details).

Lastly, the questions had to be phrased in a way as not to make the decision completely wrong or right because gender discrimination would not be as evident in a question from which the respondents' choices were blatantly obvious. For example, if we had offered a question on whether a Sailor should be discharged from the U.S. Navy for a sexual assault conviction, our presumption is that the majority if not all respondents, regardless of gender, would indicate that the Sailor should indeed be discharged.

1. What's in a Name?

Names selected for the survey questions had to be distinctly recognizable as male or female. To ensure that respondents were not unsure of the gender associated with the particular question they were reviewing, an examination of the most popular male and female baby names was conducted via the U.S. Social Security Administration (SSA) website (2018). The SSA collects name data via state agencies as a part of their requirement for social security number assignment. Male names selected were William and Jacob, and female names selected were Emily and Mary, as each of these names were in the top 130 names since 1990 (SSA, 2018). None of the names selected was in the top 1,000 most popular names for the opposite gender during this same period.

.

2. Gender Questions

Both the promotion recommendation and end-of-tour award questions were assessed on a Likert scale from 1 (completely inappropriate) to 7 (completely appropriate). Appropriateness of promotion recommendation and award assigned were based on the following scenarios.

a. Promotion Recommendation Question

The importance of FITREPs is established early and often in a naval officer's career. A report of one's individual performance, the FITREP not only impacts a sailor's career, it also signals to the Navy who it can rely on to achieve its current and future missions. Due to the critical importance of FITREPs, we believe that a scenario depicting a male or female officer receiving a questionable FITREP will engage the participant and undoubtedly lead some to make impassioned choices regarding appropriateness of said recommendation. To that end, our first survey question portrays a male or female Navy lieutenant who has just received a "must promote" (MP) performance recommendation on the individual's most current FITREP. The scenario presents a situation in which the officer has seen a reduction in promotion category for missing recruiting mission more times than the Navy average in the individual's second year, after meeting or exceeding mission each month in the first year. The number of times the individual missed mission was set at four versus the Navy average missing mission three times in an attempt to keep the responses toward the middle of the Likert scale.

For our research, the most important detail is the gender-specific name and pronouns throughout the scenario as this variable may elicit conscious or unconscious biases as to whether the mark given, given a specific gender, is more or less appropriate than for someone of the opposite gender. Concern for impassioned responses toward one end of the Likert scale or another is mitigated by the randomization of the survey questions and ultimately the data will not be affected as we are only concerned with measuring the difference

between gender pairs. That is, the data we seek is not whether the performance appraisal (FITREP) system is fair or unfair, but rather whether there are any statistically significant differences between genders in the rewarding of a particular promotion recommendation.

As FITREP content and "requirements" for performance vary widely across officer communities the difficulty with this type of question was to provide enough information for the participant to make an informed decision without requiring him or her to have specific knowledge of any one community. For this reason, we choose to create a scenario of a U.S. Navy recruiter, a line of work that most if not all officers have some general knowledge of, experience in, or interactions with (see Appendix C for question content).

b. End-of-Tour Award Question

Our second survey question portrays a male or female Navy lieutenant who has just received an "end-of-tour" award for his or her performance as a division officer. Like promotions, a personal award is a topic that all naval officers should be familiar and experienced with. Yet again, the difficulty with this type of question is providing enough material to allow the respondent to answer, while at the same time not leading the responder down a path that requires more extensive knowledge of the situation. The back story for the question is that of a lieutenant who has received a Navy Commendation Medal (NCM) for a job that had historically only warranted a Navy Achievement Medal (NAM). Respondents were given supporting information such as an examination of the award justification; a review of her peers' FITREPs in which they performed "marginally" better; and a note establishing that although they did perform "marginally" better, they had inherited a "seasoned and well trained" support staff. The reasoning for these comments was to justify the warranting of the reward but to also indicate that they may not have warranted an increase in award precedence from a NAM to an NCM. Once more, this was an attempt to keep the question neutral. As with FITREPs, the researchers acknowledge that passions run high when talking

about awards, and who does or does not warrant them. Again, concern for impassioned responses toward one end of the Likert scale or another are mitigated by the randomization of the survey questions, and ultimately, the data will not be affected as we are only concerned with measuring the difference between gender pairs. That is, the data we seek is not whether the awards system is fair or unfair, but whether there are any statistically significant differences between the genders in the rewarding of an award (see Appendix C for question content).

3. Demographic Questions

Demographic questions asked in our survey include participants rank, age, gender, marital status, and NPS School. Although multiple demographic questions were asked, the only variable of interest is gender. Participant gender allows us to measure the difference between their responses and the gender of the fictitious officer in the question, resulting in the following gender pairs: male-male, male-female, female-male, and female-female. The other demographic data allows us to look at the make-up of our survey population, while serving as distractors so as not to clue the participant in to the true intent of the survey.

4. Pilot Survey

In order to improve the readability and effectiveness of the questions posed, a pilot survey was distributed to the March 2018, NPS Graduate School of Business and Public Policy, Masters of Science in Management, 847 Curriculum, graduating class. Data obtained from 21 participants was used to improve the study instrument. The data from the pilot was not included in our final analyses.

The trial promotion recommendation question had two variants with group assignment executed by the random function in Microsoft Excel (<.5 & >.5). The first question set (delta 3), has the Navy average missing mission three times, and the officer missing mission six times. The second question set (delta 1), has

the Navy average missing mission three times, the officer missing mission four times. Gender did not change as we were seeking to test the effect of the difference between the Navy meeting mission and the officer missing mission would have on responses. The end-of-tour award question remained unchanged.

Analysis of the pilot indicated that on average, females rate marginally higher across both versions (agree more so with the reduction in promotion recommendation), of the promotion recommendation question. Furthermore, females, on average, marginally rate lower across both versions of the end-of-tour award question (disagree more so with the awarding of the higher award).

The averages for all answers were between 3.33 and 4.67. Interestingly the largest change in answers between the two groups was not for the delta 3 and delta 6 questions; it was for the end-of-tour award question (average of 4 and 2.125, respectively). Extreme answers included one participant rating each question as a 7, and another rating the questions at 0.

Final analysis of the data and comments resulted in a subtle refinement of the information contained in both questions, as well as a decision to include the delta 1 version of the promotion recommendation question as the average sat more in the middle of the Likert scale (see Appendix E for pilot survey questions and comments).

E. SURVEY DISTRIBUTION

Active duty Navy NPS students were sent an email invitation to participate in the survey via LimeSurvey, an on-line survey distribution and data collection tool. As the NPS student population consists of all service branches, we restricted the distribution to the Navy population as other branches may not be familiar enough with Navy FITREPs and Navy personal award requirements. We further restricted ourselves to active duty officers at NPS to complete the research within a constrained timeframe. Consent was obtained from those who chose to take part in the survey. The consent informed the students that they were invited to participate in a Navy survey on the “Perceptions of Reward and

Punishment in the U.S. Navy." After the survey had officially closed, the students were then sent a follow-on email that acted as a secondary consent in which the true intent of the research was identified (see Appendix D for follow-on email). The initial consent and the follow-on email were approved by the NPS IRB as a way of ensuring that participants were not harmed, while protecting the integrity of the research process (see Appendix A for email invitation).

F. PARTICIPANTS

The sample population of participants was limited to current resident student naval officers at NPS. The population contained a variety of school departments and demographics. Emails were sent to 585 students at NPS with 264 entering the survey for a response rate of 45 percent. Of those who entered the survey, 235 completed the survey for a completion rate of 89 percent. One participant requested that his data be excluded and removed from the dataset after receiving the secondary consent and decreased our final dataset to 234 complete responses.

G. DEMOGRAPHICS

Our survey population demographics are described in Table 2. Here, we discover that our participant population is 82 percent male with 87 percent being between the rank of Lieutenant (O3) and Lieutenant Commander (O4). We also see that our population's age is concentrated between the ages of 25 and 44 (92.6 percent), with 62 percent of the population being married. The participants are fairly evenly distributed among three school houses (GSBPP, GSEAS, and GSIOS), containing nearly 86 percent of our population.

Table 2. Survey population demographics

Descriptor	Obs	% of Pop.
Gender		
Male	192	82
Female	42	18
Military Rank (by paygrade)		
O1	4	2
O2	14	6
O3	145	62
O4	58	25
O5	8	3
O6	0	0
Rank other	5	2
Age		
< 25	6	3
25-29	64	27
30-34	77	33
35-39	62	26
40-44	20	9
45-49	5	2
Marital Status		
Single	89	38
Married	145	62
NPS (affiliated school)		
GSBPP	59	25
GSEAS	75	32
GSIOS	67	29
GSIOS	5	2
NPS other	28	12

H. VARIABLES

Our dataset contains variables that fall into two broad categories: Participant Gender and Survey Question Variables, with variables representing dichotomous, ordinal and nominal values.

a. Participant Variables

As described in Chapter IV, Section C.3, “Demographic Questions,” the only participant variable of interest is gender.

b. Survey Question Variables

The variables in our survey questions are the questions themselves as rated by the participant on a Likert scale. The participants are offered a choice of seven values from 1 (completely inappropriate) to 7 (completely appropriate), with the neutral point being 4 (appropriate). This format allows the participant to express how much the participant agrees or disagrees with the scenario and its outcome.

Summary statistics for survey question 1 (Female/Male Recruiter Scenario) and question 2 (Male/Female Nurse Scenario) are shown in Table 3. Question 1 and question 2 are packaged so that the participants were randomly distributed a group of questions that were matched male recruiter/male nurse or female recruiter/female nurse. Statistics from our recruiter question are that the version containing a female recruiter was completed by 118 (96 male/ 22 female) participants with a mean score of ~3.12. For the second version of the question containing a male recruiter, we see that 116 (96 male/20 female) received this version for a mean score of ~ 3.03. Statistics from our nurse question are that the version containing a female recruiter was completed by 118 (96 male/ 22 female) participants with a mean score of ~2.93. For the second version of the question containing a male nurse, we see that 116 (96 male/20 female) participants received this version for a mean score of ~ 3.12. Differences in means across gender pairs are discussed in our findings.

Table 3. Survey question summary statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
FemRec	118	3.12	1.46	1	7
MaleRec	116	3.03	1.56	1	7
FemRN	118	2.93	1.27	1	7
MaleRN	116	3.12	1.52	1	7

I. OLS REGRESSION

With our variables defined and their summary statistics shown, we now test for differences in population means by participant gender, by gender survey question. We evaluate the gender "punishment" and "reward" questions independently as Likert scale ratings for these questions have different interpretations. In performing this analysis we employ OLS regression to predict the value of the survey question based on the value of the independent variable (male). To conduct this analysis, we employ the following specification:

$$Q_i = \beta_0 + \beta_1 Male + \varepsilon$$

where:

Q_i is the randomly assigned survery question

$Male$ is the respondent's gender, 1 if male, 0 otherwise

The results of the mean differences for the fictitious officer recruiter question (punishment) are shown in Table 4, while the results for the fictitious officer nurse question (reward) are shown Table 5.

Table 4. Participant gender by recruiter question

	Female Participant			Male Participant		
	Obs	Mean	Std. Dev.	Obs	Mean	Std. Dev.
FemRec	22	3.14	1.25	96	3.11	1.51
MaleRec	20	2.75	2.75	96	3.09	1.61
Diff		0.39 (I)			0.02 (II)	

I. Female participants rate males lower (Female*FemRec-Female*MaleRec) when compared within same gender.

II. Male participants rate males lower (Male*FemRec-Male*MaleRec) when compared within same gender.

Note. Lower ratings on the recruiter question signify that the rater deems it more inappropriate for the fictitious officer recruiter to receive a decrease in promotion recommendation.

For our recruiter question, in which we asked the appropriateness of a decrease in promotion recommendation for a male or a female officer recruiter, lower ratings signify that the rater deems it more inappropriate for the fictitious officer recruiter to receive a decrease in promotion recommendation. For interpretation by the female participant, we find that that on average females possess a positive bias toward males (0.39), or negative bias toward females. As for male participants, we find that there is no discernable difference in the means, as the difference (-0.01) is less than 2 percent of the standard deviation. In summary, we find that females have a small but measurable negative bias toward females. However, the applied regression model cannot statistically significantly predict the dependent variable, FemRec; F (1, 115) = 0.02, p = 0.8795, or MaleRec; F (1, 115) = 0.81, p = 0.3711.

Table 5. Participant gender by nurse question

	Female Participant			Male Participant		
	Obs	Mean	Std. Dev	Obs	Mean	Std. Dev.
FemRN	22	2.86	1.08	95	2.95	1.32
MaleRN	20	2.75	1.55	97	3.20	1.51
Diff		0.11 (I)			-0.25 (II)	

I. Female participants rate males lower (Female*FemRN-Female-MaleRN) when compared within same gender.

II. Male participants rate females lower (Male*FemRN-Female*MaleRN) when compared within the same gender.

Note. Lower ratings on the nurse question signify that the rater deems it more inappropriate for the fictitious nurse officer to receive an increase in award precedence.

For our nurse question, in which we asked about the appropriateness of a perceived increase in award precedence from a NAM to an NCM for a male or female nurse officer, lower ratings signify that the rater deems it more inappropriate for the fictitious nurse officer to receive an increase in award precedence. For interpretation by the female participant, we find that that on average females possess a negative bias toward males (0.11), or a positive bias

toward females. As for male participants, we find that on average males possess a negative bias toward females (-0.25), or a positive bias toward males. In summary, we find that both females and males have a small but measurable negative bias toward the opposite gender. Nonetheless, the applied regression model cannot statistically significantly predict the dependent variable, FemRN; $F(1, 115) = 0.08$, $p = 0.7821$, or MaleRN; $F(1, 115) = 1.51$, $p = 0.2224$.

J. CHAPTER SUMMARY

This chapter provided the context of our survey question design, development and distribution. As our focus was on the identification of gender bias in the U.S. military, a field experiment methodology such as an audit or correspondence study seemed an appropriate approach to test whether gender bias exists within the active duty U.S. Navy officer population at NPS.

This chapter also presented the data collected, methods of analysis, and the corresponding results. Our results indicate that there is a detectable, yet subtle bias, in our recruiter (perceived punishment) question as female participants displayed a negative bias (more apt to agree to a decrease in promotion recommendation) toward females. Of greater interest, our findings indicate the presence of gender bias, mainly favoring matched gender pairs in our nurse (perceived reward) question. That is, we find that both females and males have a small but measurable negative bias toward the opposite gender. The model applied, however, could not statistically significantly predict the applied dependent variables. Although the observed differences that we found could be real, it is more likely that the study suffered from being underpowered or that the observed differences occurred simply due to chance.

THIS PAGE INTENTIONALLY LEFT BLANK

V. DATA, METHODS, AND RESULTS: STUDENT EVALUATIONS OF TEACHING

A. INTRODUCTION

This chapter builds on the findings of our review of the literature on student evaluations of teaching and discusses the data, methods, and results of our analysis. We first provide a framework for the NPS SOFs and NPS environment, as well as the benefits they offer our study. We next introduce our rich and uniquely advantaged dataset before discussing our variables, and their summary statistics. Having provided context to our environment and data, we turn to answering our research questions by discussing the methods used to uncover the existence of gender bias in our population, and the subsequent results. We employ several methods of analysis to compare evaluations by student and instructor gender pairs. We begin with *t* tests to determine whether any statistically significant differences emerge in mean responses. We then employ models for the evaluation score of 1–5, and Linear Probability Models (LPM), using student and course fixed effects, and instructor and course fixed effects, for each of the SOF questions. These models further help identify gender bias by controlling for numerous variables. The use of multiple methods serves to validate our findings.

B. FRAMEWORK

The SOF consists of 15 questions, which students assess online using a five-point Likert scale, as shown in Table 6. Of greatest importance to this study is question 12 (Q12) as this question is meant as a summative measure of overall instructor effectiveness and carries the most weight in decisions on instructor careers. The SOF is open to students one week prior to the completion of the relevant quarter of study. If students do not complete their SOF, they will not receive a grade for that class.

Table 6. NPS SOF questions

Naval Postgraduate School Student Opinion Form Questions	
Q1	The course was well organized.
Q2	Time in class was spent effectively.
Q3	The instructor seemed to know when students didn't understand the material.
Q4	Difficult concepts were made understandable.
Q5	I had confidence in the instructor's knowledge of the subject.
Q6	I felt free to ask questions.
Q7	The instructor was prepared for class.
Q8	The instructor's objectives for the course have been made clear.
Q9	The instructor made the course a worthwhile learning experience.
Q10	The instructor stimulated my interest in the subject area.
Q11	The instructor cared about student progress and did his/her share in helping others to learn.
Q12	Overall, I would rate this instructor:
Q13	Overall, I would rate this course:
Q14	Overall, I would rate the textbook(s):
Q15	Overall, I would rate the quality of the exams:

Numerous aspects of the NPS environment offer our analysis an advantage over other studies. This advantage stems from the unique characteristics of NPS: military students attending a military school. Class attendance is compulsory, which helps reduce some of the bias caused by an inadequate basis to judge, as discussed in Chapter III, Part D. NPS students, through compulsory attendance, have a greater exposure to their instructors and, therefore, are in a better position to more accurately assess instructors. While attendance alone will not fully inform students to accurately judge teacher effectiveness, it does better inform the NPS population over those at other institutions where attendance may not be compulsory.

SOF completion is also compulsory at NPS and has a 99.1 percent completion rate (Arkes & Eger, personal communication, November 2017). As every student must complete a SOF, the sample is free from non-response bias. We also avoid self-selection bias with students randomly allocated to instructors—a strength of other studies (Boring, 2017; Carrell et al., 2010; Hoffman & Oreopoulos, 2007). Although students can request to shift classes to another time, or with another instructor, these requests are uncommon and

heavily scrutinized to limit any second order effects, mainly the maintenance of equal class sizes.

Another unique aspect of NPS is its student population. By far, most SET research involves populations of a much younger age at undergraduate institutions. The younger the audience, the less informed their evaluations are. NPS is a postgraduate school offering Master's and doctoral degrees. Student ages at NPS typically range from 25 to 39 compared to the typical undergraduate age range of 18 to 22. NPS students also have strong military backgrounds, evidenced by the competitive selection process to attend the school. The military experience of NPS students typically ranges from five to 15 years. These differences allow us to determine more stable results in our analysis, through a sample of experienced, mature, and professional military officers.

C. DATASET

Our dataset is the same dataset used, and was provided by Jeremy Arkes and Robert Eger (personal communication, November, 2017). The dataset originated from NPS' Institutional Research, Reporting, and Analysis office that is responsible for the collection and reporting of SOF data. The raw data file contained 363,000 student-level SOF rating observations, over the period FY2007–FY2016.

Jeremy Arkes and Robert Eger restricted their dataset to enhance the integrity of their analysis (personal communication, November, 2017). First, they restricted the dataset by removing any observations not assigned a letter grade. Omitting these observations removes students who have completed a "pass/fail subject," as well as students who may have withdrawn from the class. By maintaining only those classes assigning a letter grade to students allows us to control for grades; an important variable as discussed throughout our literature review (Alauddin & Kifle, 2014; Wright & Jenkins-Guarnieri, 2012; Addison et al., 2006; Comm & Mathaisel, 1998).

Further, Jeremy Arkes and Robert Eger (personal communication, November, 2017) removed observations on class sizes below seven, as well as above 50 students. The logic for this restriction is that directed study classes, as opposed to regular classes, fall below seven students, and classes above 50 students are not typical. Given the influence of class size on evaluation scores (Bedard and Kuhn, 2006), these restrictions ensure that the effect of class size in our analysis is representative of the sample population. We however, restricted upper class size to 35 students, as classes above 35 are greater than two standard deviations from the class size mean. Finally, they removed observations without associated student or faculty identifiers, to maintain the integrity of observations.

D. VARIABLES

Our dataset contains variables that we group into three categories: dependent, key explanatory, and control variables, with variable types representing: discrete, dichotomous, and ordered polychotomous values. The types of data that NPS collects allows us to take advantage of the findings in other research regarding variable use, as discussed in Chapter III. Specifically, we have data on student and instructor gender, academic discipline, class size, grade, and academic rank. As we have variables for both student and instructor gender we are also able to generate variables for the four student and instructor gender interactions, forming our key explanatory variables. These interactions give us an advantage over other research, which due to the anonymous nature of evaluations does not often collect data on student gender.

1. Dependent Variables

The development of our econometric model requires a dependent variable whose value is determined by the model's right-hand side variables. Our analysis uses the SOF rating, for each of the SOF's 15 questions, as our dependent variables. Using each SOF question as a dependent variable allows us to determine how each of our right-hand side variables influence a change in our

dependent variable. Questions 1–15 are discrete variables, taking a value of 1 through 5, representing the evaluation given by a student to an instructor for a given class. We also analyze the probability of an instructor achieving a rating of 5 versus a rating of 1–4, represented as dichotomous dependent variables, allowing the use of LPMs. The summary statistics for our dependent variables are shown in Table 7, which shows us that Q14 (textbook rating) has the lowest mean value of 3.89 and Q5 (confidence in instructor knowledge) has the highest mean value of 4.71. We also find that an instructor has the greatest probability of obtaining a rating of 5 for Q5 (confidence in instructor's knowledge) at 78.3 percent, and Q14 (textbook rating) has the lowest probability of obtaining a rating of 5 at 36.7 percent.

Table 7. Dependent variable summary statistics

Variable	Obs	Mean	Std. Dev.	$P(Q_i = 5)$
q1	173,741	4.36	0.89	0.562
q2	172,317	4.27	0.96	0.529
q3	172,584	4.24	0.98	0.520
q4	172,553	4.24	0.95	0.502
q5	173,652	4.71	0.63	0.783
q6	173,622	4.64	0.72	0.742
q7	172,831	4.59	0.76	0.709
q8	173,506	4.41	0.89	0.611
q9	173,418	4.34	0.96	0.586
q10	172,961	4.26	1.02	0.551
q11	172,262	4.49	0.83	0.653
q12	173,723	4.32	0.94	0.562
q13	173,617	4.12	0.99	0.441
q14	142,522	3.89	1.11	0.367
q15	154,815	4.12	1.00	0.497

2. Key Explanatory Variables

As we aim to test for gender bias in how students evaluate instructors through the SOF, we are interested to see whether differences exist between student and instructor gender pairs. To do this we generate interactions for the four student and instructor pairs, which form our key explanatory variables for our

econometric analysis. These variables are male student with male instructor (MSMI), male student with female instructor (MSFI), female student with male instructor (FSMI), and female student and female instructor (FSFI). Each of these interactions is dichotomous variables taking a value of 1 if the gender pair is met, and a 0 otherwise. The summary statistics for our student instructor gender interactions are shown in Table 8. We find our data comprises 88 percent male students (MS), 82 percent male instructors (MI), 72 percent MSMI, 16 percent MSFI, 10 percent FSMI, and 2 percent FSFI.

Table 8. Key explanatory variable summary statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Male student	175,093	0.88	0.33	0	1
Male instructor	175,093	0.82	0.38	0	1
Male student & Male instructor	175,093	0.72	0.45	0	1
Male student & Female instructor	175,093	0.16	0.36	0	1
Female student & Male instructor	175,093	0.10	0.30	0	1
Female student & Female instructor	175,093	0.02	0.15	0	1

3. Control Variables

We have grouped all variables that we can to control for as "control variables," which represent the variables deemed as key influences of evaluation scores from Chapter III. These variables include: the student's grade for the class (GPA); the instructor's experience (Experience); whether the instructor is tenure tracked (a proxy for academic rank); the relevant school to which the evaluation relates—Graduate School of Business and Public Policy (GSBPP), Graduate School of Engineering and Applied Sciences (GSEAS), Graduate School of Operational and Information Sciences (GSOIS), and School of International Graduate Studies (SIGS); and, class size. The GPA variable is a discrete variable, representing the student's letter grade for the class being evaluated, as a numeric value according to a four-point GPA scale (A = 4, A-=3.7, B+=3.3, etc.). Experience is a discrete variable representing how many years an

instructor has taught at NPS. Tenure track is a dichotomous indicator variable. The school variables: GSBPP, GSEAS, GSOIS, and SIGS, are also dichotomous indicator variables. Class size is a discrete variable representing the number of students in a class. The control variable summary statistics are described in Table 9, which informs us: the mean level of instructor experience is 11.22 years; 53 percent of instructors are tenure tracked; GSBPP represents 27 percent of our sample population; GSEAS 30 percent; GSIOS 30 percent; SIGS 13 percent; and Class size has a mean value of 19.31.

Table 9. Control variable summary statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
GPA	174,954	3.71	0.41	1	4
Experience	175,093	11.22	8.28	0	53
Tenure track	175,093	0.53	0.50	0	1
GSB	175,093	0.27	0.44	0	1
GSEAS	175,093	0.30	0.46	0	1
GSOIS	175,093	0.30	0.46	0	1
SIGS	175,093	0.13	0.34	0	1
Class size	175,093	19.18	6.65	7	35

4. NPS Gender Representation

The gender representation of both student and instructor for each NPS school is shown in Table 10. This is relevant as it provides further context to the gender balance within our dataset. To provide these summary statistics, we generated interactions of student and instructor genders with school variables. GSBPP has the highest female student (FS) population (15 percent); SIGS has the highest female instructor (FI) population; GSOIS has the highest MS population (31 percent); and both GSEAS and GSOIS have the highest MI population (87 percent).

Table 10. Gender representation summary statistics

Variable	Obs	Mean	Std. Dev	Min	Max
For GSBPP:					
Male student	47,074	0.85	0.36	0.00	1.00
Male Instructor	47,074	0.77	0.42	0.00	1.00
For GSEAS:					
Male student	52,985	0.87	0.33	0.00	1.00
Male Instructor	52,985	0.87	0.33	0.00	1.00
For GSOIS:					
Male student	52,089	0.90	0.29	0.00	1.00
Male Instructor	52,089	0.87	0.34	0.00	1.00
For SIGS:					
Male student	22,945	0.88	0.32	0.00	1.00
Male Instructor	22,945	0.69	0.46	0.00	1.00

E. *T* TESTS OF STUDENT AND INSTRUCTOR GENDER PAIRS

With our variables defined and their summary statistics shown, we now test for differences in population means by student and instructor gender pairs, by SOF question. To enable this analysis, we generated interactions for each of the four gender pairs: MSMI, MSFI, FSMI, and FSFI. We conducted two-tailed *t* tests to test the following hypothesis for each SOF question:

$$H_0: \mu_1 = \mu_2$$

$$H_A: \mu_1 \neq \mu_2$$

The results from the *t* tests of student and instructor gender pairs by SOF question are show in Table 11. We find that on average: male students provide higher ratings to instructors than do female students, and that female instructors receive higher ratings from students than do male instructors. These results hint at possible gender bias favoring female instructors, particularly by male students; however, the differences in mean scores are small in scale and do not account for other variables that may influence results.

Table 11. *T* tests of student and instructor gender pairs by SOF question

	Male instructor		Female instructor		
	Obs	Mean	Obs	Mean	*Diff*
Q1					
Male student	125,521	4.353	27,080	4.403	-0.050***
Female student	17,152	4.317	3,988	4.398	-0.081**
Diff		0.036***		0.005**	
Q2					
Male student	124,516	4.272	26,923	4.288	-0.016***
Female student	16,924	4.204	3,953	4.270	-0.066**
Diff		0.068***		0.018**	
Q3					
Male student	124,725	4.234	26,941	4.345	-0.111***
Female student	16,959	4.111	3,959	4.276	-0.165**
Diff		0.123***		0.069**	
Q4					
Male student	124,696	4.234	26,863	4.326	-0.092***
Female student	17,035	4.136	3,959	4.285	-0.149**
Diff		0.098***		0.041**	
Q5					
Male student	125,460	4.721	27,060	4.691	0.030***
Female student	17,146	4.679	3,986	4.685	-0.006**
Diff		0.042***		0.006**	
Q6					
Male student	125,433	4.641	27,062	4.701	-0.060***
Female student	17,136	4.539	3,991	4.639	-0.100**
Diff		0.102***		0.062**	
Q7					
Male student	124,864	4.590	26,957	4.597	-0.007***
Female student	17,031	4.563	3,979	4.602	-0.039**
Diff		0.027***		-0.005**	
Q8					
Male student	125,343	4.403	27,043	4.466	-0.063***
Female student	17,132	4.394	3,988	4.477	-0.083**
Diff		0.009***		-0.011**	
Q9					
Male student	125,273	4.342	27,039	4.387	-0.045***
Female student	17,120	4.270	3,986	4.367	-0.097**
Diff		0.072***		0.020**	

Q10					
Male student	124,987	4.263	26,956	4.301	-0.038***
Female student	17,041	4.152	3,977	4.277	-0.125**
Diff		0.111***		0.024**	
Q11					
Male student	124,476	4.485	26,847	4.574	-0.089***
Female student	16,986	4.424	3,953	4.542	-0.118**
Diff		0.061***		0.032**	
Q12					
Male student	125,498	4.324	27,080	4.354	-0.030***
Female student	17,151	4.254	3,994	4.343	-0.089**
Diff		0.070***		0.011**	
Q13					
Male student	125,423	4.119	27,065	4.154	-0.035***
Female student	17,138	4.042	3,991	4.141	-0.099**
Diff		0.077***		0.013**	
Q14					
Male student	102,152	3.883	23,917	3.960	-0.077***
Female student	13,036	3.793	3,417	3.919	-0.126**
Diff		0.090**		0.041**	
Q15					
Male student	112,564	4.116	24,025	4.190	-0.074***
Female student	14,798	4.017	3,428	4.095	-0.078**
Diff		0.099***		0.095***	

Although these results appear to provide evidence of gender bias, they do not account for other variables that may affect results in either direction. Given these limitations when using *t* tests for assessing multidimensional problems, we must employ further techniques to uncover gender bias and its influencing factors.

F. ECONOMETRIC MODELS

When evaluating instructor effectiveness, we employ two different econometric models, to test for the presence of gender bias among our population. First, we evaluate the correlation that Q1–11 have with Q12, independently, by gender pairs. Finally, we use two OLS models: one with the dependent variable as the actual evaluation (ranging 1–5), and the other with the

dependent variable indicating whether the evaluation was a “5” or not. We employ student and course fixed effects, and instructor and course fixed effects to analyze all SOF questions to see the effect of our key explanatory variables (student and instructor gender pairs) on SOF ratings. The use of multiple methods allows us to cross examine results and validate findings.

1. Independent SOF Question Correlations with Question 12

Our first model analyzes what SOF questions can be determined as the best predictors of instructor effectiveness. As Q12 (student’s overall rating of the instructor) is the main measure that NPS uses to assess instructor effectiveness, an analysis of how other SOF questions can predict the rating of Q12 (student’s overall rating of the instructor) is a valid approach, particularly when interrogating differences in gender pairs.

Using Q12 (student’s overall rating of the instructor) as the dependent variable, representing the student’s perception of instructor effectiveness, we regress Q1–11 (Q13–15 are omitted as we feel they are too subjective), in turn, to compare the coefficient of determination value (R-squared) across gender pairs. If our population is free from gender bias, we expect no major differences across gender pairs in R-squared values. This approach stemmed from a validation method used by Boring (2017). She analyzed the differences in how students evaluated teachers when grouping SET questions by the dimensions of teaching, which she linked to gender stereotypes. We were unable to ascertain how NPS align SOF questions to dimensions of teaching, so we are unable to adopt a similar approach; however, we assess the correlations that each question has on Q12 (student’s overall rating of the instructor), as an effective means of further interrogating for the presence of gender bias and somewhat akin to Boring’s approach. Despite the importance placed on understanding the implications of gender stereotypes throughout our literature review, when seeking to test for gender bias, our approach uses correlations to sort the most predictive questions of effectiveness, not the application of semi-subjective teaching

dimensions and gender stereotypes. To conduct this analysis, we employ the following OLS specification:

$$Q12GenderPair_{i,j,c} = \beta_1 Q_{1\to 11} + \varepsilon_{i,j,c}$$

where:

$Q12GenderPair_{i,j,c}$ is the gender of student i and instructor j in class c

Q is the SOF question being regressed on $Q12$ ranging from 1– 11

The results of the R-squared values of Q1–11, regressed on Q12 (student's overall rating of the instructor), by student and instructor gender pairs, ranked from high to low are shown in Table 12. This table shows little difference in how each of the student and instructor gender pairs order the importance of SOF questions as predictors for Q12. Noting how close values of R-squared are, and how a small change in R-squared may have a large change in ranking, these results provide further evidence in support of the finding in table 18: there is little indication of gender bias in NPS student evaluations of instructors, using this method.

Table 12. R-squared values of questions 1–11 regressed on Q12

Male student & male instructor		Male student & female instructor		Female student & male instructor		Female student & female instructor	
Q	R-squared	Q	R-squared	Q	R-squared	Q	R-squared
9	0.689	9	0.684	9	0.694	9	0.698
10	0.609	10	0.606	10	0.621	10	0.608
2	0.601	2	0.585	2	0.592	4	0.587
4	0.586	4	0.568	4	0.591	2	0.575
3	0.577	3	0.564	11	0.582	3	0.568
11	0.561	1	0.545	3	0.573	11	0.532
1	0.545	7	0.505	1	0.536	1	0.528
8	0.515	8	0.504	8	0.502	7	0.514
7	0.488	11	0.5	7	0.486	8	0.51
5	0.361	5	0.425	6	0.395	5	0.457
6	0.355	6	0.314	5	0.383	6	0.396

Ranking from high to low R-squared value

2. Fixed Effects Models

Our next approach is to employ more sophisticated econometric models to analyze the influence of student and instructor gender pairs on SOF ratings when controlling for several student, instructor, class and school variables. Our rich dataset allows us to take advantage of models using student and course fixed effects, and models using instructor and course fixed effects. As fixed effects models use within-group estimates, they remove the influence of across-class invariant characteristics; both observed and unobserved thereby, greatly mitigating omitted variable bias. As time-invariant characteristics drop out of the model, we avoid the complexity of otherwise having to measure, or omitting, difficult traits (effectiveness, motivation, innate ability, difficulty, etc.). The only limitation is that fixed effects models do not correct for bias introduced from omitted variables that vary across class.

3. Student and Course Fixed Effects

Employing student and course fixed effects estimates the weighted average of within-group (student) coefficient estimates—or, how students rate female instructors compared to male instructors when also factoring out time invariant course effects. The use of student fixed effects, while having advantages, does not account for the possible confounding factor of actual differences in quality and effectiveness between male and female instructors.

4. Instructor and Course Fixed Effects

Models with instructor and course fixed effects estimate the weighted average of within group (instructor) coefficient estimates— or, how instructors are rated by male students compared to female students when factoring out time invariant course effects. When using instructor fixed effects, we do not account for the confounding factor that there may be actual differences in how male and female students evaluate their instructors.

5. Econometric Model Specifications

The student and course fixed effects models use the following specification:

$$SOF_{i,j,c} = GenderPair_{i,j,c}\beta_1 + X\beta_2 + +\mu_i + \mu_c + \varepsilon_{i,j,c}$$

where:

$SOF_{i,j,c}$ is how student i evaluates instructor j in class c for each SOF question

$GenderPair_{i,j,c}$ is the gendered pair between student i and instructor j in class c

X is a set of control variables

μ_i is a set of student fixed effects

μ_c is a set of course fixed effects

The instructor and course fixed effects models use the below specification:

$$SOF_{i,j,c} = GenderPair_{i,j,c}\beta_1 + X\beta_2 + \mu_j + \mu_c + \varepsilon_{i,j,c}$$

where:

μ_j is a set of instructor fixed effects

We estimate these models using OLS and the two forms of dependent variables described earlier: the actual evaluation (ranging from 1–5) and an indicator for whether the evaluation was a "5." For the later dependent variable, to be consistent across models, we use an LPM. We employ an LPM over the use of an ordered logistic regression, probit or logistic regression to avoid what would otherwise be a comparison to an awkward mean individual. This comparison is well explained by Williams (2012) where the average individual of comparison was someone "who is 47.57 years old, 10.5 percent black and 52.5 percent female" (p. 324) when discussing marginal effects at the means.

The use of our two OLS models using student and course fixed effects, and instructor and course fixed effects provides us with two perspectives assisting us understand the effects of gender bias across our models. Therefore, coefficient estimates must be interpreted in the context of their model design.

6. Evaluation Score of 1 to 5 Fixed Effects Models and Results

To estimate the impacts of student and instructor gender on evaluation score of 1–5 we first use two fixed effects models, one employing student and course fixed effects and the other employing instructor and course fixed effects. The results of our fixed effects models, controlling for student, instructor, class, and school variables, for SOF questions 1–15, are shown numerically as Table 13 and visually as Figure 2 for the student and course fixed effects model (shown as a percentage of a standard deviation (SD)), and Figure 3 for the instructor and course fixed effects model (shown as a percentage of a SD).

Table 13. Fixed effects models for the evaluation score of 1 to 5

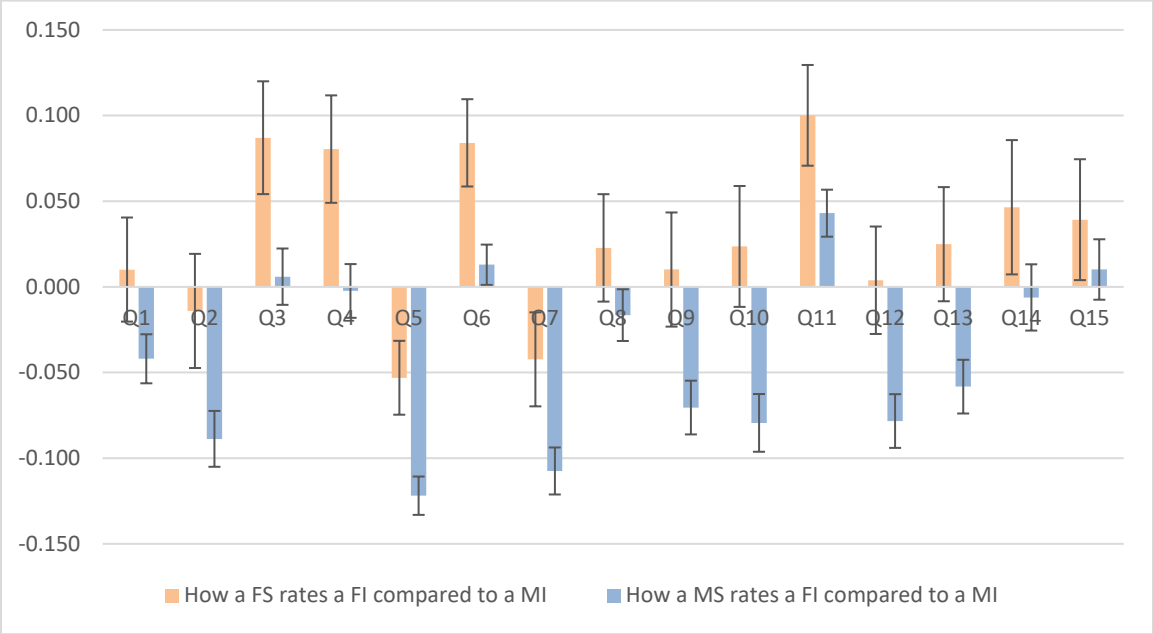
Dependent variable	(1) FSFI *(ref: FSMI)*	(2) MSFI *(ref: MSMI)*	(3) MSFI *(ref: FSFI)*	(4) MSMI *(ref: FSMI)*
Q1 - The course was well organized	0.009	-0.037***	-0.031**	0.025***
	(0.016)	(0.007)	(0.014)	(0.007)
Q2 - Time in class was spent effectively	-0.014	-0.085***	-0.019	0.058***
	(0.017)	(0.008)	(0.015)	(0.007)
Q3 - The instructor seemed to know when students didn't understand the material	0.085***	0.006	0.027*	0.108***
	(0.017)	(0.008)	(0.015)	(0.007)
Q4 - Difficult concepts were made understandable	0.076***	-0.002	-0.007	0.081***
	(0.016)	(0.008)	(0.015)	(0.007)
Q5 - I had confidence in the instructor's knowledge of the subject	-0.034***	-0.077***	-0.018*	0.032***
	(0.011)	(0.006)	(0.010)	(0.005)
Q6 - I felt free to ask questions	0.060***	0.009	0.043***	0.096***
	(0.013)	(0.006)	(0.012)	(0.006)
Q7 - The instructor was well prepared for class	-0.032**	-0.081***	-0.042***	0.015***
	(0.014)	(0.007)	(0.012)	(0.006)
Q8 - The instructor's objectives for the course have been made clear	0.020	-0.015*	-0.041***	0.000
	(0.016)	(0.008)	(0.014)	(0.007)
Q9 - The instructor made the course a worthwhile learning experience	0.010	-0.068***	-0.017	0.060***
	(0.017)	(0.008)	(0.015)	(0.007)
Q10 - The instructor stimulated my interest in the subject area	0.024	-0.081***	-0.018	0.092***
	(0.018)	(0.009)	(0.016)	(0.008)
Q11 - The instructor cared about student progress	0.083***	0.036***	0.007	0.055***
	(0.015)	(0.007)	(0.013)	(0.006)
Q12 - Overall, I would rate this instructor	0.004	-0.073***	-0.031**	0.059***
	(0.016)	(0.008)	(0.015)	(0.007)
Q13 - Overall, I would rate the course	0.025	-0.058***	-0.026*	0.059***
	(0.017)	(0.008)	(0.016)	(0.007)
Q14 - Overall, I would rate the textbook(s)	0.051**	-0.007	-0.004	0.053***
	(0.020)	(0.010)	(0.019)	(0.010)
Q15 - Overall, I would rate the quality of exams	0.039**	0.010	0.031*	0.062***
	(0.018)	(0.009)	(0.017)	(0.008)
Student FE	Yes	Yes	No	No
Instructor FE	No	No	Yes	Yes
Course FE	Yes	Yes	Yes	Yes

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

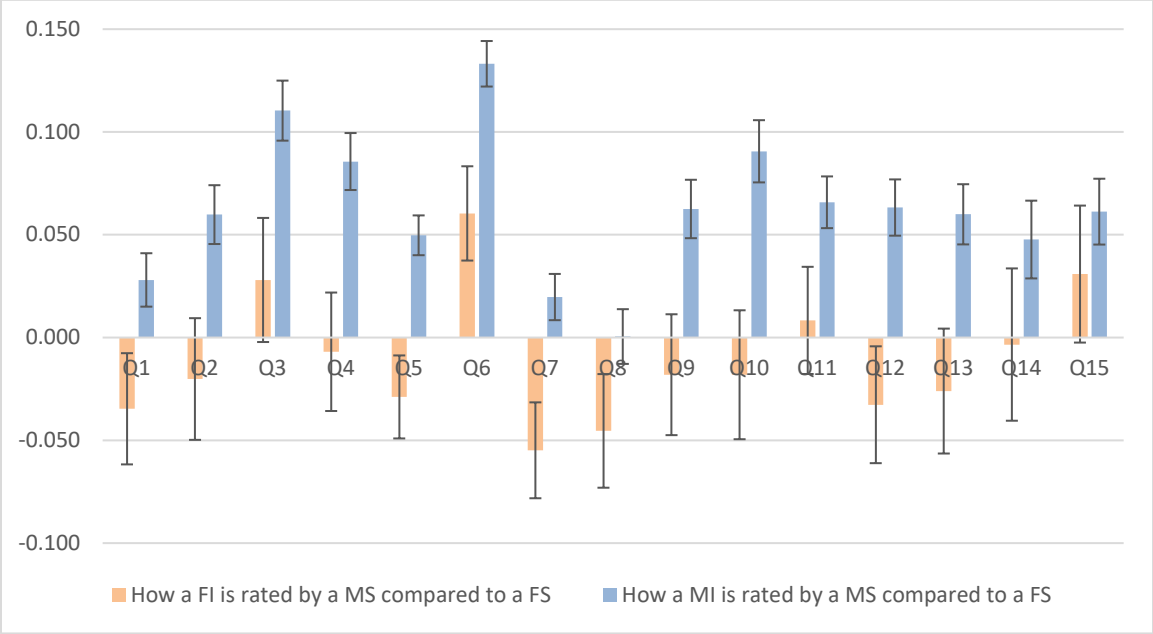
Coefficient estimates on student/instructor gender pairs controlling for student, instructor, class and school variables, for SOF questions 1–15.

Figure 2. Evaluation score of 1 to 5 student and course fixed effects model gender pair vectors



y axis = percentage of a SD, 95% CI shown.

Figure 3. Evaluation score of 1 to 5 instructor and course fixed effects model gender pair vectors



y axis = percentage of a SD, 95% CI shown.

Table 13 and Figures 2 and 3 inform us that both models show a favor to matched gender pairs, suggesting students show gender bias in favor of a same sex instructor. These results also show that the favor towards matched gender pairs is strongest among male students. From our student and course fixed effects model, column 1 of Table 13, we find female students favor female instructors, when compared to male instructors, for six SOF questions with statistical significance. Here, the largest magnitude concerns Q3 (understanding student's grasp of material) with a coefficient estimate of 0.085 (8.7 percent of a SD). Whereas female students favor male instructors for only two questions: Q5 (confidence in instructor knowledge) and Q7 (instructor preparation), with coefficient estimates of -0.034 and -0.032, respectively (5.3 and 4.2 percent of a SD). We find, when looking at Q12 (student's overall rating of the instructor) that female students show little difference toward instructor gender with a coefficient estimate of 0.004 (0.4 percent of a SD) in favor of female instructors, which is statistically insignificant. As there are differences in how females evaluate male and female instructors, we are provided with insights into how female students value instructors, given instructor gender. Hereafter, we will only describe the estimates, for evaluation score of 1–5 models, in terms of SD.

In contrast, of our statistically significant results, male students favor male instructors for nine SOF questions, column 2 Table 13. Here, the largest magnitudes are for Q5 (confidence in instructor's knowledge), and Q7 (class preparation) with magnitudes of 12.2, and 10.7 percent of a SD, respectively. For Q12 (student's overall rating of the instructor) male students favor male instructors by 7.8 percent of a SD, statistically significant. As with our female students, these results inform us that male students value different attributes from male and female instructors however, with a much stronger preference for male instructors.

What is also interesting is where favor is demonstrated by both student genders to the same instructor gender, columns 1 and 2 of Table 13. Our student and course fixed effects models show us that male and female students rate

male instructors higher than female instructors with statistically significant results for Q5 (confidence in instructor's knowledge), and Q7 (instructor preparation), with male students providing higher ratings. This finding may suggest students are rating based off perceived gender and role stereotypes with the effect strongest within the matched gender pair. Similarly, female instructors are rated higher than male instructors for Q11 (caring about student progress), by both student genders, with female students giving higher ratings. This finding may also suggest students evaluate toward perceived gender and role stereotypes with the effect again strongest within the matched gender pair.

Our instructor and course fixed effects model shows statistically significant results, for female students rating female instructors higher than male students, for six SOF questions, column 3 Table 13. Of these questions, Q7 (instructor preparation) and Q8 (clear objectives) have the highest magnitudes of 5.5 and 4.5 percent of a SD. Female instructors are rated higher by male students, compared to female students, with statistical significance for three SOF questions. These questions are Q3 (understanding student's grasp of material), Q6 (free to ask questions), and Q15 (exam quality) with magnitudes of 2.8, 6.0, and 3.1 percent of a SD, respectively. These results inform us that male and female students value different qualities from a female instructor. For Q12 (student's overall rating of the instructor) we find female students evaluate female instructors higher than male students by 3.3 percent of a SD.

When we look at the differences in how female and male students evaluate male instructors, column 4 Table 13, we find that male students rate male instructors higher than female students for all SOF questions that have statistically significant results. The largest magnitudes are for Q3 (understanding student's grasp of material), Q6 (free to ask questions), and Q10 (stimulated interest) with magnitudes of 11.0, 13.3, and 9.1 percent of a SD, respectively. These findings could suggest that male students are showing a positive bias to male instructors, or female students are showing a negative bias to male instructors, or any combination in between—reflected purely by the evaluations

of male instructors. Q12 (student's overall rating of the instructor) when evaluated by a male student is 6.3 percent of a SD higher than when evaluated by a female student.

Our fixed effects models for the evaluation score of 1–5 demonstrate that male students provide higher evaluation scores than female students, particularly in favor of male instructors. We also find male and female students show favor towards the same sex instructor, particularly higher among male students. We also uncover questions that appear to be influenced by perceived gender and role stereotype characteristics. We learn that male students value different instructor characteristics than female students, and that this varies dependent on the instructor's gender. When we look specifically at Q12 (student's overall rating of the instructor) we find that students favor matched gender pairs where the effect is much larger within the male matched pair. As NPS is highly populated with male students and male instructors, the higher ratings given by male students, and the favor shown by male students towards male instructors, disadvantages female instructors.

7. LPM Fixed Effects Models and Results

We now employ our LPM using student and course fixed effects, and instructor and course fixed effects. Our LPM informs us how our gender pairs influence the predicted probabilities of an instructor receiving an evaluation score of 5, compared to a 1–4 when controlling for our student, instructor, class, and school characteristics. This is an advantageous approach, when looking to identify gender bias, as results indicate differences in achieving the highest evaluation score. If gender bias is present then we would expect to see large and significant differences across our gender pairs, as individuals would be less likely to rate those they are biased against at the highest level. The results of our LPM fixed effects models, for SOF questions 1–15, are shown numerically as Table 14 and visually as Figure 4 for the student and course fixed effects model, and Figure 5 for the instructor and course fixed effects model.

Table 14. LPM fixed effects models

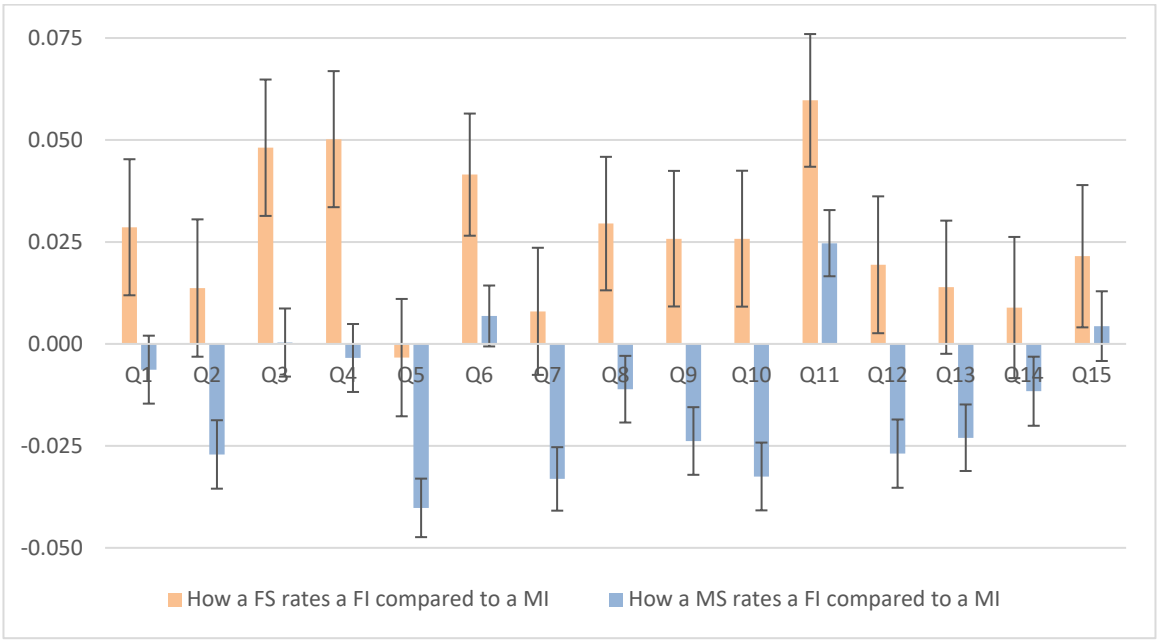
Dependent variable	(1) FSFI *(ref: FSMI)*	(2) MSFI *(ref: MSMI)*	(3) MSFI *(ref: FSFI)*	(4) MSMI *(ref: FSMI)*
Q1 - The course was well organized	0.029***	-0.006	-0.025***	0.010***
	(0.009)	(0.004)	(0.008)	(0.004)
Q2 - Time in class was spent effectively	0.014	-0.027***	-0.014*	0.027***
	(0.009)	(0.004)	(0.008)	(0.004)
Q3 - The instructor seemed to know when students didn't understand the material	0.048***	0.000	-0.013	0.038***
	(0.009)	(0.004)	(0.008)	(0.004)
Q4 - Difficult concepts were made understandable	0.050***	-0.003	-0.024***	0.028***
	(0.009)	(0.004)	(0.008)	(0.004)
Q5 - I had confidence in the instructor's knowledge of the subject	-0.003	-0.040***	-0.018***	0.024***
	(0.007)	(0.004)	(0.007)	(0.003)
Q6 - I felt free to ask questions	0.042***	0.007*	0.018**	0.053***
	(0.008)	(0.004)	(0.007)	(0.003)
Q7 - The instructor was well prepared for class	0.008	-0.033***	-0.033***	0.011***
	(0.008)	(0.004)	(0.007)	(0.004)
Q8 - The instructor's objectives for the course have been made clear	0.030***	-0.011***	-0.045***	-0.005
	(0.008)	(0.004)	(0.008)	(0.004)
Q9 - The instructor made the course a worthwhile learning experience	0.026***	-0.024***	-0.020**	0.029***
	(0.008)	(0.004)	(0.008)	(0.004)
Q10 - The instructor stimulated my interest in the subject area	0.026***	-0.033***	-0.020**	0.040***
	(0.009)	(0.004)	(0.008)	(0.004)
Q11 - The instructor cared about student progress	0.060***	0.025***	-0.013*	0.020***
	(0.008)	(0.004)	(0.008)	(0.004)
Q12 - Overall, I would rate this instructor	0.019**	-0.027***	-0.023***	0.030***
	(0.009)	(0.004)	(0.008)	(0.004)
Q13 - Overall, I would rate the course	0.014*	-0.023***	-0.015*	0.026***
	(0.008)	(0.004)	(0.008)	(0.004)
Q14 - Overall, I would rate the textbook(s)	0.009	-0.012***	-0.002	0.021***
	(0.009)	(0.004)	(0.009)	(0.004)
Q15 - Overall, I would rate the quality of exams	0.022**	0.004	0.011	0.028***
	(0.009)	(0.004)	(0.009)	(0.004)
Student FE	Yes	Yes	No	No
Instructor FE	No	No	Yes	Yes
Course FE	Yes	Yes	Yes	Yes

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

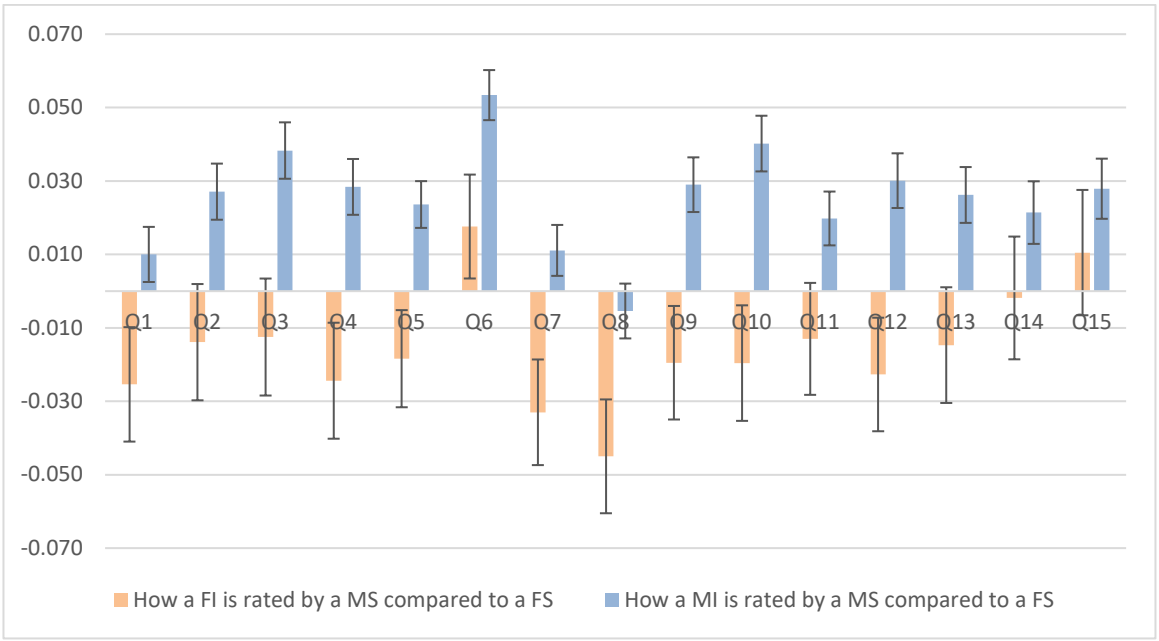
Coefficient estimates on student/instructor gender pairs controlling for student, instructor, class and school variables, for SOF questions 1–15.

Figure 4. LPM student and course fixed effects model gender pair vectors



y axis = percentage points, 95% CI shown

Figure 5. LPM instructor and course fixed effects model gender pair vectors



y axis = percentage points, 95% CI shown

Table 14 and Figures 4 and 5, echo findings from our fixed effects models for the evaluation score of 1–5 by showing stronger student favor towards matched gender pairs. Our student and course fixed effects model shows that female instructors are more likely to be evaluated a 5 by female students than male students, for all questions with statistical significance (11 SOF questions), column 1 Table 14. The largest likelihood that female students would evaluate female instructors with a 5, compared to male students, is for Q11 (cared about student progress) by 6.0 percentage points, with statistical significance. For Q12 (student's overall rating of the instructor) female students are more likely to evaluate female instructor with a 5 than male instructors by 1.9 percentage points and is statistically significant. These findings show that female students place a higher value on female instructors than male instructors.

Similarly, male students are more likely to evaluate male instructors with a 5 than female instructors for nine SOF questions, column 3 Table 14; the highest difference in likelihood being Q5 (instructor knowledge) at 4.0 percentage points, statistically significant. Males students however, are more likely to evaluate female instructors with a 5, with statistical significance, for Q6 (felt free to ask questions) by 0.7 percentage points, and Q11 (cared about student progress) by 2.5 percentage points. Q12 (student's overall rating of the instructor) has male students more likely to evaluate male instructors with a 5 than female instructors by 2.7 percentage points and is statistically significant. Mirroring the results from our fixed effects models for the evaluation score of 1–5, we learn that there are differences in what male students value from instructors and they are dependent on the instructor's gender.

Our student and course fixed effects model shows student gender agreement in favor of a female instructor, column 1 and 2 of Table 14, for Q6 (free to ask questions), and Q11 (instructor cared about student progress). As with our fixed effects models for the evaluation score of 1–5, these findings show students rating, as we would expect, if evaluations are influenced by perceived

gender and role stereotypes. These finding are again strongest within the matched gender pair.

Our instructor and course fixed effects model informs us that male students are more likely to evaluate female instructors with a 5 than female students, column 3 of Table 14, for only one SOF question, Q6 (free to ask questions) with statistical significance. Here, male students are more likely to rate female instructors with a 5 than female students by 1.8 percentage points with statistical significance. Female instructors are more likely to be evaluated a 5 by female students for 11 SOF questions with statistical significance. Of these 11 questions, when compared to male students, Q8 (clear objectives) has the greatest magnitude of 4.5 percentage points, closely followed by Q7 (instructor preparation) at 3.3 percentage points. Q12 (student's overall rating of the instructor) has a value of 2.3 percentage points seeing female students more likely to evaluate female instructors with a 5 than male students. As we find differences between how female instructors are rated by male and female students, we again learn that male and female students value differing qualities from female instructors.

This model also informs us that male students are consistently more likely to evaluate male instructors a 5 than female students, column 4 of Table 14, for 14 SOF questions, with statistical significance. Here, the largest magnitudes occur for Q3 (understanding student's grasp of material) at 3.8 percentage points, Q6 (free to ask questions) at 5.3 percentage points, and Q10 (stimulated interest) at 4.0 percentage points. Q12 (student's overall rating of the instructor) shows that male students are more likely to rate male instructors a 5, by 3.0 percentage points, than female students. These results align to our instructor and course fixed effects models for the evaluation score of 1–5, which could suggest that male students are showing a positive bias to male instructors, or female students are showing a negative bias to male instructors, or any combination in between—reflected purely by differences in the likelihood that a male instructor is evaluated a 5.

Our LPM models align to the findings of our fixed effects models for the evaluation score of 1–5, with stronger results. We find that students favor matched gender pairs; however, they agree on instructor gender preference for questions that align more closely to traditionally perceived gender and role stereotype characteristics. We also find that male students are more generous when evaluating male instructors. Although female students are also more likely to evaluate female instructors a 5, the effect is not as large when compared to male matched gender pairs. We also uncover that male and female students value different instructor characteristics, and that this varies with the instructor's gender. Given the dominance of male representation at NPS (88 percent male students and 82 percent male instructors), this finding puts male instructors at an advantage over female instructors.

G. CONCLUSION

This chapter introduced the framework for our analysis and described the rich dataset available for our analysis, identifying the advantages of our population in cleanly uncovering gender bias over other like research. We then conducted *t* tests of our student and instructor gender pairs and identified that male students provide higher instructor ratings than female students do and that female instructors receive higher ratings than male instructors; however, these results do not account for other variables that may affect results. We then looked at the correlations that each SOF Q1–11 has on Q12, independently, by gender pair, to see if instructor effectiveness is measured differently by each gender pair. Our results for this approach did not find any differences in how effectiveness is measured across our gender pairs. Finally, we employed OLS models for the evaluation score of 1–5, and for an evaluation score of 5, using student and course fixed effects, and instructor and course fixed effects, to further scrutinize our population for signs of gender bias.

Our findings suggest the presence of gender bias, mainly favoring matched gender pairs, where the effect is largest among male matched pairs.

Due to the heavily weighted representation of male students at NPS, any preference for gender matched pairs puts male instructors at an advantage over their female peers in student evaluations. We also find evidence of student ratings aligning, in terms of instructor gender, to questions more clearly associated with traditional perceived gender and role stereotype characteristics. This indicates that while stereotypes appear to have an influence on evaluations, it is not evident across all questions. Therefore, military gender equality training and a greater exposure of military members to females (through the opening of all military job specialties to females) may be diminishing, or positively altering, the negative effects of gender and role stereotypes. As the results, of student gender alignment towards an instructor gender, differ in magnitude between student genders

Our OLS models for the evaluation score of 1–5, and for an evaluation score of 5, using instructor and course fixed effects, showed questions where both male and female students show the same preference towards an instructor gender. Although there is alignment in same-gender instructor preference, the magnitude of the effect differs by student gender. While this may indicate evidence consistent with the presence of a positive gender bias by the student gender showing the greatest favor, it may in fact be a negative bias being shown by the other student gender, or any combination in between. Where the true value lies remains unknown however, is most likely in the middle.

As discussed at the start of Chapter V, our student and course fixed effects, and our instructor and course fixed effects models provide different points of view and are both subject to confounding factors: that there may be actual differences between how male and female students evaluate male and female instructors, and there may also be differences in actual male and female instructor effectiveness. Uncovering differences in what male and female students value from their instructors, and finding that this varies by instructor gender, may be reflected by our inability to control for these confounding factors.

VI. CONCLUSION

A. INTRODUCTION

This thesis examined whether gender bias exists within the U.S. military. We first introduced the background to our research with Chapter I, to provide the reader with a cognitive framework to appreciate the significant impacts of gender bias at organizational and individual levels. We then presented two literature reviews, Chapters II and III, one for each of our analytical methods (our survey and our SOF analysis), which provided us with insights to the contemporary findings of literature and the various methods employed by researches to uncover gender bias. We then provided our data, methods and results from our survey analysis, Chapter IV, and for our SOF analysis, Chapter V.

This chapter concludes our research by revisiting and answering our primary and secondary research questions before making recommendations for future research, as well as touching on methods for mitigating gender bias.

B. RESEARCH QUESTIONS

Having compared the results for our two analytical methods we now revisit our primary and secondary research question to provide answers.

1. Primary Research Questions

Our first primary research question was: ***Does gender bias exist within the U.S. military?***

The results of our survey indicate that there are measurable differences in how females and males respond to the questions posed. The differences were most evident in our reward scenario in which differences in ratings across our four gender pairs are indicative of a negative bias toward the opposite gender. Nevertheless, the model failed to deliver statistically significant differences, and a conclusive finding of gender bias in the U.S. military cannot be determined at this time via our survey.

Our SOF analysis uncovered evidence consistent with gender bias within our population. Our findings indicated differences in evaluation scores across our four gender pairs. This determination was consistent throughout our analysis, which leads us to conclude that evidence suggests gender bias does exist within the U.S. military. However, alternate explanations may also account for these differences—there may be actual differences between male and female instructors, and there may be actual differences in how male and female students evaluate instructors.

Our second primary research question was: ***Does student gender demonstrate a bias towards the gender of an individual they are evaluating?***

Results from our survey parallel the findings of our first research question. While our recruiter (punishment) question indicates a negative bias against females by females, our nurse (reward) question reveals a negative bias against the opposite gender by both females and males. As with question 1 these findings are in light of our models' failure to deliver statistically significant results.

Our SOF analysis found that both male and female students tend to favor instructors of the same sex. We also found that favoritism in gender matched pairs is larger within male matched gender pairs. Our SOF analysis also uncovered that when male and female students both favor an instructor of a specific gender, it is for those questions that adhere closely to traditionally perceived gender and role stereotype characteristics. These findings were consistent across the various econometric models that we employed, and we therefore conclude that student and instructor gender may influence the level of bias shown.

Our final primary research question was: ***If gender bias does exist, how prevalent is it?***

Findings of our survey cannot adequately demonstrate the magnitude or direction of gender bias; therefore, it would be inappropriate to comment on its

prevalence in our population even though the coefficient estimates in our model suggest subtle bias.

The prevalence of bias is revealed by the magnitude of coefficient estimates in our econometric models. Considering Q12 (student's overall rating of the instructor), our student and course fixed effects model for the evaluation of score of 1–5 showed that male students rate male instructors higher than female instructors by 7.8 percent of a SD, whereas female students did not have a statistically significant result when evaluating instructor effectiveness. Our instructor and course fixed effects model for the evaluation score of 1–5 showed female students rate female instructors higher than male students by 6.3 percent of a SD and male students rate male instructors higher than female students by 3.3 percent of a SD.

Our LPM for Q12 (student's overall rating of the instructor) reveals the same preference toward matched gender pairs. We found, from our student and course fixed effects model, that female student rate female instructors higher than male instructors by 1.9 percentage points, and that male students evaluate male instructors higher than female students by 2.7 percentage points. Similar findings are reflected in our instructor and course fixed effects model where female students rate female instructors higher than male students by 2.3 percentage points and male students rate male instructors higher than female students by 3.0 percentage points.

We find that while our models reveal student favor towards matched gender pairs, the size of these effects are small overall, which suggests a marginal prevalence of gender bias when evaluating instructor effectiveness. However, due to NPS having a large representation of male students and male instructors, a preference for male matched gender pairs places female instructors at a disadvantage. Furthermore, we are estimating average effects, so some instructors may be more affected than others by any gender bias. We also must note that our models do not account for any actual differences between male and female instructors, or any actual differences between how male and female

students evaluate instructors. As we are unable to control for this confounding factor, our results may be capturing this effect biasing our estimates away from true casual estimates.

2. Secondary Research Questions

Our first secondary research question was: ***How do we measure gender bias?***

Measuring gender bias is a difficult task. Due to the multifaceted and complex human decision-making process, the ability to cleanly identify and measure what part, if any, of a decision is attributable purely to gender bias is arguably impossible. However, we employed two methods that sought to minimize the impacts of numerous other factors, allowing us to attribute gender bias to the differences in how our gender pairs responded to our survey and within the SOFs.

Our survey analysis was able to employ techniques found in audit and correspondence studies. These field experiment methodologies have been instrumental in the detection of bias in the housing and labor markets. Although the observed differences that we found could be real, it is just as likely that the observed differences occurred simply due to chance. With that being said, the literature substantiates the aforesaid field experiment methodologies as a means of capturing and measuring gender bias.

Our SOF analysis was able to control for those variables that literature has found to be influential on evaluation scores. Furthermore, our unique population ensured our analysis was free of self-selection and non-response bias, as well as mitigating for a student's inadequate basis to judge. These advantages have allowed us to measure gender bias within our population and in the context of the population environment.

Our final secondary research question was: ***What methods could be used to mitigate gender bias?***

We suggest two approaches for mitigating gender bias and its effects on decision making. First, we recommend an ongoing policy of detection, which could be achieved through the employment of anonymous surveys throughout the organization. Collection and interrogation of results would determine whether gender bias exists, and where, leading to targeted programs to educate participants on the detrimental organizational and individual effects of gender bias.

We also recommend the known use of “dummy” resumes for selection boards and other similar selection activities. We believe that if selection boards are aware that some of the presented enlisted or officers are “fake,” then the members of the board will be forced to think more critically about their decision, which would greatly reduce the impacts of unconscious decision influences. Further, consideration should be given to “blind” resumes and biographies for selection boards. Implementing such a measure may yield significant gains in selections of quality, as was found in Goldin and Rouse (2000).

While our findings do indicate the presence of gender bias, the magnitudes of bias are marginal. This indicates that, in our population, the military is proving to be an organization that treats males and females fairly equal. This is a positive discovery that indicates the positive nature of military training on appreciating equality and may also indicate the positive step in opening all military specialties to females and thus providing a greater exposure to diversity, which can diminish the effects of stereotypes within the organization.

C. FUTURE RESEARCH

Due to the limitations discussed in Chapter I, the use of further research into gender bias in the U.S. military would go a long way to further substantiate our findings. Concerning our survey analysis, the results lead us to recommendations that may enhance future studies utilizing our gender bias survey methodology.

Our first recommendation for future studies in our survey would be to consider the use of gender pronouns only versus gender-specific names and pronouns to reduce likelihood of revealing the survey's true objective. This recommendation is in light of a few participants' comments that they viewed the questions as more apt to capture gender bias, rather than their view on rewards and punishment. Of the 264 participants who entered the survey, one individual commented that the activity was likely a "gender bias hunt," while another commented: "Why are both fictional characters in this study women? That seems like it would inherently skew the responses of the survey taker based on experiences with women receiving awards and the performance evaluation system." However, with this last statement we feel as though the intended audience does realize that there are certain attitudes and feelings that are developed by their "experiences with women receiving awards."

Our second recommendation for the use of a survey to capture bias would be to ask or design questions that are relatable to Sailors in their specific communities. For example, one of our respondents commented: "What is important for each of these communities? What do they care about? What is important? How is it different from my community? I really don't feel qualified to answer how appropriate it is for any of these people to be promoted or awarded because of how the Navy structures these things for each community. I have no idea if that is good, bad, or average for each of these sailors for what they do." Another respondent added that the fictitious officers "are both performing in supporting roles within the Navy construct instead of fleet work in the Unrestricted Line. Many takers of the survey might not be able to understand the requirements placed upon those women to be successful in their positions." For this reason, a recommendation for targeted research within communities may produce a higher survey completion rate as those who failed to complete the survey may have felt as if they could not answer the questions based on their experiences and the details provided.

Future research concerning our SOF analysis should follow a similar approach but target other military schools, across enlisted and officer ranks, with inclusion of the student's military specialty as an additional categorical variable. This will test whether the results of this thesis are generalizable across the wider military community and determine whether levels of bias differ across ranks and by job specialties. Another area of scope would be to include military instructors, in addition to civilian instructors. Doing so will test whether our findings are generalizable to military instructors and could reveal whether military students hold military instructors to different standards, which we consider plausible

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. RECRUITING EMAIL FOR PARTICIPANTS

Dear Recipient,

As a resident USN Active Duty Naval Postgraduate School **(NPS)** student, staff, or faculty, you are invited to participate in a **5 minute** survey that will benefit a research study titled **"Perceptions on Reward and Punishment within the U.S. Navy."**

This research study is a degree requirement for the Master of Science in Management degree at the NPS and is being conducted by Luke Siwek and Brandon Wolf. This survey addresses a variety of career scenarios that you and your Sailors may find yourselves in. Navy leadership relies on the information from this survey to evaluate the effectiveness of existing policies and determine where changes are needed.

You can access the survey by clicking on the survey link provided below. After reading the online consent form on LimeSurvey, by clicking on the "Yes" button, you are acknowledging that you have read and understand this information and that you agree to voluntarily participate in this research survey.

Questions about your rights as a research participant or any other concerns may be addressed to the NPS, IRB Chair, Dr. Larry Shattuck, 831–656-2473, lgshattu@nps.edu. If you have any questions or comments about the research, please contact the Principal Investigator, Dr. Jeremy Arkes, 831–656-3819, jaarkesc@nps.edu.

IP addresses and personally identifiable information (PII) will be collected. Any information that is obtained during this study will be kept confidential to the full extent permitted by law. All efforts, within reason, will be made to keep your personal information in your research record confidential but total confidentiality cannot be guaranteed. All PII will be destroyed once the research data has been collected.

To participate, please click on this link: https://survey.nps.edu/341311/lang-en
Very Respectfully,

Luke Siwek
lsiwek@nps.edu

Brandon Wolf
bwolf@nps.edu

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B. INITIAL SURVEY CONSENT

Perceptions on Reward and Punishment within the U.S. Navy

Naval Postgraduate School
Consent to Participate in Research

Introduction. You are invited to participate in a research study titled "Perceptions on Reward and Punishment within the U.S. Navy." The purpose of the research is to gain insight on how Naval officers view rewards and punishments within the U.S. Navy. This is a research is being conducted by thesis students at the Naval Postgraduate School in Monterey, CA.

Procedures. Participants will enter their responses via an on-line survey that will take approximately 5 minutes to complete. The researchers expect less than a 1000 subjects will be participating in the research. No compensation will be given for participation in this research.

Location. The survey will take place via LimeSurvey.

Voluntary Nature of the Study. Your participation in this study is strictly voluntary. If you choose to participate you can change your mind at any time and withdraw from the study. You will not be penalized in any way or lose any benefits to which you would otherwise be entitled if you choose not to participate in this study or to withdraw.

Potential Risks and Discomforts. The potential risk is a breach of confidentiality.

Anticipated Benefits. Anticipated benefits from this study include the identification and mitigation of unfair reward and punishment systems. Additionally, decisions and judgments of individuals in supervisory roles may be enhanced by this research. You may not directly benefit from your participation in this research.

Confidentiality & Privacy Act. Any information that is obtained during this study will be kept confidential to the full extent permitted by law. All efforts, within reason, will be made to keep your personal information in your research record confidential but total confidentiality cannot be guaranteed. Your responses will be confidential; however, identifying information such as your name, email address, and IP address wil be collected. All data is stored in a password protected electronic format. The results of this study will be used for scholarly purposes only and may be shared with Naval Postgraduate School

representatives. Once we have completed our data collection all PII (name and email) will be destroyed.

Points of Contact. If you have any questions or comments about the research, or you experience an injury or have questions about any discomforts that you experience while taking part in this study please contact the Principal Investigator, Jeremy Arkes, 831–656-3819, 2646, jaarkesc@nps.edu. Questions about your rights as a research subject or any other concerns may be addressed to the Navy Postgraduate School IRB Chair, Dr. Larry Shattuck, 831–656-2473, lgshattu@nps.edu.

Statement of Consent. I have read the information provided above. I have been given the opportunity to ask questions and all the questions have been answered to my satisfaction. I have been provided a copy of this form for my records and I agree to participate in this study. I understand that by agreeing to participate in this research and signing this form, I do not waive any of my legal rights.

Select **"Yes"** if you wish to participate in the research.

Select **"No"** if you do not consent to participate in the research.

Please choose only one of the following:

- ○ Yes
- ○ No

APPENDIX C. SURVEY QUESTIONNAIRE

Perceptions on Reward and Punishment within the U.S. Navy

Rank *
Please choose only one of the following:
- ○ O-1
- ○ O-2
- ○ O-3
- ○ O-4
- ○ O-5
- ○ O-6
- ○ other

Age *
Please choose only one of the following:
- ○ <25
- ○ 25-29
- ○ 30-34
- ○ 35-39
- ○ 40-44
- ○ 45-49
- ○ > 49

Gender *
Please choose only one of the following:
- ○ Male
- ○ Female
- ○ other

Marital Status *
Please choose only one of the following:
- ○ Single
- ○ Married

NPS School (if you are not affiliated with a listed NPS School, please select "other") *
Please choose only one of the following:
- ○ GSBPP
- ○ GSEAS
- ○ GSOIS
- ○ SIGS
- ○ other

LT Mary Jones is a Navy recruiter. In her first year of recruiting she met or exceeded mission each month, as too did the Navy. Subsequently, LT Jones received a promotion recommendation of "early promote." The next year, while the Navy failed to meet its monthly mission 3 times, she failed to meet mission 4

times. LT Jones' promotion recommendation after that year was "must promote." There were no changes in LT Jones' reporting senior or promotion status from year one to year two. *

Only answer this question if the following conditions are met: ((random == 1))

Please choose the appropriate response for each item:

	1 Completely Inappropriate	2	3	4 Appropriate	5	6	7 Completely Appropriate
Based on the information given and in your opinion, is LT Jones' performance recommendation appropriate?	○	○	○	○	○	○	○

LT William Jones is a Navy recruiter. In his first year of recruiting he met or exceeded mission each month, as too did the Navy. Subsequently, LT Jones received a promotion recommendation of "early promote." The next year, while the Navy failed to meet its monthly mission 3 times, he failed to meet mission 4 times. LT Jones' promotion recommendation after that year was "must promote." There were no changes in LT Jones' reporting senior or promotion status from year one to year two. *

Only answer this question if the following conditions are met: ((random == 2))

Please choose the appropriate response for each item:

	1 Completely Inappropriate	2	3	4 Appropriate	5	6	7 Completely Appropriate
Based on the information given and in your opinion, is LT Jones' performance recommendation appropriate?	○	○	○	○	○	○	○

LT Emily Smith, a Nurse Corps Officer, has just received the Navy Commendation Medal (NCM) for her "end of tour" award while serving as the Division Officer of a Naval Clinic. Historically, LT's that have held this position and successfully completed their tour have been awarded the Navy Achievement Medal (NAM), which is of lower precedence than the NCM. Upon review of the award's justification, as well as her peers' Fitness Reports, it appears that LT Smith performed her duties marginally better than her peers, noting she had inherited a "seasoned and well trained" support staff. *

Only answer this question if the following conditions are met: ((random == 1))

Please choose the appropriate response for each item:

	1 Completely Inappropriate			**4 Appropriate**			**7 Completely Appropriate**
Based on the information given and in your opinion, is the level of award that LT Smith received appropriate?	○	○	○	○	○	○	○

LT Jacob Smith, a Nurse Corps Officer, has just received the Navy Commendation Medal (NCM) for his "end of tour" award while serving as the Division Officer of a Naval Clinic. Historically LT's that have held this position and successfully completed their tour have been awarded the Navy Achievement Medal (NAM), which is of lower precedence than the NCM. Upon review of the award's justification, as well as his peers' Fitness Reports, it appears that LT Smith performed his duties marginally better than his peers, noting he had inherited a "seasoned and well trained" support staff. *

Only answer this question if the following conditions are met: ((random == 2))

Please choose the appropriate response for each item:

	1 **Completely Inappropriate**	2	3	4 **Appropriate**	5	6	7 **Completely Appropriate**
Based on the information given and in your opinion, is the level of award that LT Smith received appropriate?	○	○	○	○	○	○	○

Comments or Suggestions?

Please write your answer here:

Thank you for your participation in this important research!

Submit your survey.
Thank you for completing this survey.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX D. FOLLOW-ON EMAIL

Follow-On Email (notification of deception)

Dear Recipient,

You were recently invited to participate in a research study titled “Perceptions on Reward and Punishment within the U.S. Navy.”

This email is to inform you that the true intent of the research study was to identify the existence of Conscious and/or Unconscious Gender Bias instead of the originally stated intent of gaining insight in to “how naval officers view rewards and punishments within the U.S. Navy.”

In order to collect the data the use of “deception” was approved by the Naval Postgraduate School Institutional Review Board (IRB).

All information that was obtained during this study continues to be kept confidential to the full extent permitted by law. Once we have completed our data collection all PII (name and email) will be destroyed.

If you would like to have your data excluded from this study, please contact the Principal Investigator, Jeremy Arkes, 831–656-3819, 2646, jaarkesc@nps.edu by 21 March 2018.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX E. PILOT SURVEY: PARTICIPANT COMMENTS AND RECOMMENDATIONS FOR IMPROVEMENT

Demographic questions:
Rank: [O1, O2, O3, O4, O5, O6, other]
Age: [<25, 25-29, 30-34, 35-39, 40-45, 45-49, >50]
Gender: [M, F, other]
School: [GSBPP, GSEAS, GSOIS, SIGS,]
Marital status: [married, single]

Research questions (random assignment of gender allocation to respondents):

Question 1: LT Jones is a Navy recruiter. After meeting or exceeding mission each month of his/her first year, LT Jones received a promotion recommendation of "early promote." The next year, while the Navy failed to meet its monthly mission 3 times, LT Jones failed to meet his/her mission 4 times. LT Jones' promotion recommendation after that year was "must promote." There were no changes in LT Jones' reporting senior or promotion status from year one to year two.

Is LT Jones' performance recommendation appropriate?
Scale (0-7): 0 - Not appropriate | 7 - Completely appropriate

Peer Comments [researcher comments in red]

- Seems like she does not care about achievements anymore, it is not such good motivator - earlier promotions, if the person cannot or do not want to maintain her previous or initial level of performance.
- I'm assuming here that in year one, the Navy met mission each month. I'm also assuming that this is a fair situation where LT Jones isn't assigned to a more difficult or easier recruiting assignment. I'm also not sure if "appropriate" is referring to whether she should have been given a better promotion recommendation or worse. And the scale is throwing me off because I almost want to answer this in a binary form. Appropriate or not appropriate, but I'm sure it's done in a scale for a reason, so I'll go with it! [Specify Navy met mission in year one – perhaps change final question to: In your opinion is Lt Jones' performance…].
- It depends on how much uncontrollable risk was involved during the latter recruiting season. With that said, I provide a ranking of 4, somewhat appropriate. [I feel the uncontrollable risk is implied in relation to how Navy met mission].

- EP for meeting or exceeding mission every month is good justification. Although the Navy failed to meet mission the next year 3 times, LT Jones did not meet mission one additional time. EP is typically reserved for exemplary performers, and as such, there are relatively few quotas for EP. If there was a recruiter in her command that continued to perform better than her, her MP is appropriate. [This is a fair comment, perhaps we should mention her performance in relative terms to her peers?].

- It would depend on whether the Navy failed to meet its mission the first year. If the Navy failed to meet mission and she did well despite that and got an EP, it would make sense for her to get an MP for not meeting mission. In this case I think the recommendation is completely appropriate (7). [As with a previous comment we need to mention how Navy met mission in year 1].

- If I could enter any score, I would put 3.5. If I had to assign a strict integer-only score, my response would be 3. It depends a lot on command's philosophy on whether making mission is the highest (or only) priority in assigning performance evaluation. Since I don't know how highly they value other factors like leadership, difficulty of individual environment, integrity, military bearing, etc., all I can go on is what is apparent. On that note, LT Jones exceeded mission 100 percent the first year for an EP while he or she exceeded their poor performance 4/3 times for an MP during the second period. I don't know if anything about the performance of the other recruiters, which he or she is necessarily ranked against in the MP and EP system (a relative performance system specifically). Nor do I know if their performances followed the same or different patterns relative to the national standard. All things considered, if I could enter any score, I would prefer the 3.5 on this one. [Another comment on relative performance. That said, despite front loading the questions with 'don't read into it too much,' I feel there isn't much we can do for those inclined to do so].

- I do not have enough information to determine whether the performance recommendation is appropriate or not. LT Jones could be putting in the effort but other conditions outside of LT Jones' control could be impacting mission. Also, I do not know how the Navy did during the period that LT Jones made mission and received an "early promote". Would recommend an option for the respondent that includes "Don't Know" or "Not enough information". [Again insert Navy performance for year 1. I do not want to add either 'don't know' or 'not enough information' as I imagine too many responses will pick either].

Question 2: LT Smith, a Nurse Corps Officer, has just received the Navy Commendation Medal (NCM) for his/her "end of tour" award while serving as the Division Officer of a Naval Clinic. Historically, LT's that have held this position and successfully completed their tour have been awarded the Navy Achievement Medal (NAM), which is of lower precedence than the NCM. Upon review of the award's justification, it appears that LT Smith performed her duties marginally better than his/her peers, noting he/she had inherited a "seasoned and well trained" support staff.

Is the level of award that LT Smith received appropriate?
Scale (0-7): 0 - Not appropriate | 7 - Completely appropriate

Peer Comments [researcher comments in red]

- I only gave 6, because others probably also received this award (NCM), but the person's performance was the best among others. [no action].
- Okay… I have a lot of feelings about this scenario. I never liked the concept of what a specific rank has historically been awarded, because people use that benchmark way too strictly which either underplays someone's hard work or overplays their average work. As a 2ndLt I worked my ass off at a unit and performed very well, but I was given a NAM because "2nd Lt's don't get anything higher than that." As a 1stLt at my next unit, I did fine, but nothing noteworthy and got a NCM just because other 1stLts and Capts typically got a NCM as their end of tour award. The system is very meh. [Emotive… there are rules to awards, and while I have no understanding of the U.S. military awards system, I do know that ours can sometimes spark similar emotions. However, I have found that experience does contextualize the awarding of awards. I don't think there is much we can do to satisfy this comment].
- If by noting that Lt Smith had a seasoned and well trained staff implies that her predecessors had a poor staff, then I provide a ranking of 7, completely appropriate. However, if I assume Lt Smith and her predecessors had an equally capable staff, then I would provide a ranking of 4, somewhat appropriate. [Another comment concerning relativity. However, feel we covered this by stating the award reads like she performed marginally better than her peers].
- As officers and manager, having a good staff can sometimes make an individual appear to be more talented than they are, and it is very difficult to distinguish between the two. Still not very appropriate, however. [Here I would like people to assume that the person whom wrote the award is able to appreciate the context. No action].
- I think it's appropriate. She may have been lucky but it still justifies a higher award if she did better. [Relativity, noted].

- Same response. 3.5 if that is allowed, 3 if integer only is assigned. "Marginally" better does not justify the significantly heightened award. But I have extremely incomplete information, so for all I know his or her staff was responsible. Also, I don't think it makes sense that the award write-up (justification) would note she "inherited" a good staff, this would be something that possibly comes up during the awards board, but would not be in a justification. Just a suggestion for possible rewording. [Question misread, the award write up did not say that. We may need to fix our structure to make this clearer, but clarity is relative].

- Again, I do not have enough information to determine whether the performance recommendation is appropriate or not. Maybe the individual that wrote up the justification is not a very good writer. If the award writer is known to provide accurate assessments and write them up well, I would answer with 3. If the award writer is known to provide weak assessments or write them up poorly, I would answer with 7. Either way, observers from outside the organization have to believe that the individual actually merited an NCM, otherwise the Navy award system would be meaningless. [I used to hold a belief that writing style played a part in the issuing of awards; however, at least back home, a board will sit and nominees will be discussed in detail from various points of view. If the writer is lacking in written communication, then the award will be tweaked to ensure it is suitable. As such I disagree with this evaluators POV. However, we may need to aim for greater clarity. As far as outside perceptions, well I think the issuing of awards will remain an emotive topic].

LIST OF REFERENCES

Addison, W. E., Best, J., & Warrington, J. D. (2006). Students' perceptions of course difficulty and their ratings of the instructor. *College Student Journal, 40*(2), 409–416.

Alauddin, M., & Kifle, T. (2014). Does the student evaluation of teaching instrument really measure instructors' teaching effectiveness? An econometric analysis of students' perceptions in economics courses. *Economic Analysis and Policy, 44*(2), 156–168. doi:10.1016/j.eap.2014.05.009

Arbuckle, J., & Williams, B. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles, 49*(9), 507–516. doi:1025832707002

Arrow, K. (1973). The theory of discrimination. *Discrimination in Labor Markets, 3*(10), 3–33.

Basow, S. (2000). Best and worst professors: Gender patterns in students' choices*. Sex Roles, 43*(5), 407–417. doi:1026655528055

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

Basow, S. A., Codos, S., & Martin, J. L. (2013). The effects of professors' race and gender on student evaluations and performance. *College Student Journal, 47*(2), 352–363.

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly, 30*(1), 25–35. doi:10.1111/j.1471-6402.2006.00259.x

Bedard, K., & Kuhn, P. (2008). Where class size really matters: Class size and student ratings of instructor effectiveness. *Economics of Education Review, 27*(3), 253–265. doi:10.1016/j.econedurev.2006.08.007

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

Bertrand, M., & Duflo, E. (2016). Field experiments on discrimination. NBER Working Paper Series, 22014. 10.3386/w22014

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), 991–1013.

Biernat, M., Fuegen, K., & Kobrynowicz, D. (2010). Shifting standards and the inference of incompetence: Effects of formal and informal evaluation tools. *Personality and Social Psychology Bulletin, 36*(7), 855–868. doi:10.1177/0146167210369483

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

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

Burdsal, C. A., & Harrison, P. D. (2008). Further evidence supporting the validity of both a multidimensional profile and an overall evaluation of teaching effectiveness. *Assessment & Evaluation in Higher Education, 33*(5), 567–576. doi:10.1080/02602930701699049

Carrell, S. E., Page, M. E., & West, J. E. (2010). Sex and science: How professor gender perpetuates the gender gap. *The Quarterly Journal of Economics, 125*(3), 1101–1144.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education, 71*(1), 17–33.

Comm, C. L., & Mathaisel, D. F. X. (1998). Evaluating teaching effectiveness in America's business schools: Implications for service marketers. *Journal of Professional Services Marketing, 16*(2), 163–170. doi:10.1300/J090v16n02_09

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review, 109*(3), 573–598. doi:10.1037/0033-295X.109.3.573

Fix, M., Struyk, R. (1993). Clear and convincing evidence - Measurement of discrimination in America. Washington, DC: Urban Institute Press.

Foschi, M. (2000). Double standards for competence: Theory and research. *Annual Review of Sociology, 26*(1), 21–42. doi:10.1146/annurev.soc.26.1.21

Garcia-Retamero, R., & López-Zafra, E. (2006). Prejudice against women in male-congenial environments: *Perceptions of gender role congruity in leadership. Sex Roles, 55*(1), 51–61. doi:10.1007/s11199-006-9068-1

Gillmore, G. M., & Greenwald, A. G. (1999). Using statistical adjustment to reduce biases in student ratings. *American Psychologist, 54*(7), 518–519. doi:10.1037//0003-066X.54.7.518

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *The American Economic Review, 90*(4), 715–741.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27.

Hoffman, F., & Oreopoulos, P. (2009). A professor like me: The influence of instructor gender on college achievement. *Journal of Human Resources, 44*(2), 479–494.

Kierstead, D., D'Agostino, P., & Dill, H. (1988). Sex role stereotyping of college professors. *Journal of Educational Psychology, 80*(3), 342–344. doi:10.1037/0022-0663.80.3.342

Krueger, A. B. (2002). Economic considerations and class size. NBER Working Paper 8875.

Lahey, J. N., & Beasley, R. A. (2009). Computerizing audit studies. *Journal of Economic Behavior and Organization, 70*(3), 508–514. 10.1016/j.jebo.2008.02.009

Liu, O. L. (2012). Student evaluation of instruction: In the new paradigm of distance education. *Research in Higher Education, 53*(4), 471–486. doi:10.1007/s11162-011-9236-1

MacNell, L., Driscoll, A., & Hunt, A. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education, 40*(4), 291–303. doi:10.1007/s10755-014-9313-4

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

Marsh, H. W., Bornmann, L., Mutz, R., Daniel, H., & O'mara, A. (2009). Gender effects in the peer reviews of grant proposals: comprehensive meta-analysis comparing tradition and multilevel approaches. *Review of Educational Research, 79*(3), 1290–1326.

Mau, R. R., & Opengart, R. A. (2012). Comparing ratings: In-class (paper) vs. out of class (online) student evaluations. *Higher Education Studies, 2*(3) doi:10.5539/hes.v2n3p55

Military OneSource. (2016). Demographics: Profile of the military community. Retrieved from http://download.militaryonesource.mil/12038/MOS/Reports/2016-Demographics-Report.pdf

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474. 10.1073/pnas.1211286109

Neumark, D. (2012). Detecting discrimination in audit and correspondence studies. *Journal of Human Resources, 47*(4), 1128–1157.

Neumark, D., Bank, R., & Van Nort, K. (1996). Sex discrimination in restaurant hiring: An audit study. *Quarterly Journal of Economics, cxi*(3), 915–942.

Nikolaidis, Y., & Dimitriadis, S. G. (2014). On the student evaluation of university courses and faculty members' teaching performance. *European Journal of Operational Research, 238*(1), 199–207. doi:10.1016/j.ejor.2014.03.018

Pager, D. (2007). The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future. Annals of the American Academy of Political and Social Sciences, 609 (January):104-133

Phelps, P. S. (1972). The statistical theory of racism and sexism. *The American Economic Review 62*(4), 659-661

Potvin, G., & Hazari, Z. (2016). Student evaluations of physics teachers: On the stability and persistence of gender bias. *Physical Review Physics Education Research, 12*(2), 020107. doi:10.1103/PhysRevPhysEducRes.12.020107

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education, 3*(3), 137–152. doi:10.1037/a0019865

Reisenwitz, T. H. (2016). Student evaluation of teaching: An investigation of nonresponse bias in an online context. *Journal of Marketing Education 38*(1) 7–17. doi: 10.1177/0273475315596778

Safer, A., Farmer, L. S. J., Segalla, A., & Elhoubi, A. F. (2005). Does the distance from the teacher influence student evaluations? *Educational Research Quarterly, 28*(3), 27–34.

Snowden, J. L. (2005). Explicit and implicit bias measures: Their relation and utility as predictors of criminal verdict tendency (Unpublished master's thesis). University of North Carolina, Chapel Hill, NC.

Social Security Administration. (2017). Retrieved from https://www.ssa.gov/OACT/babynames/index.html

Steinpreis, R., Anders, K., & Ritzke, D. (1999). The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. *Sex Roles: A Journal of Research, 41*(7), 509–528. 10.1023/A: 1018839203698

United States, Department of Labor, Bureau of Labor Statistics [BLS]. (2017, October 24). *Employment Projections: Civilian Labor Force, by Age, Sex, Race, and Ethnicity*. Retrieved January 6, 2018, from https://www.bls.gov/emp/ep_table_301.htm

Williams, R. (2006). Generalized ordered logit/ partial proportional odds models for ordinal dependent variables. *The Stata Journal, 6*(1), 58–82.

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education, 37*(6), 683–699. doi:10.1080/02602938.2011.563279

Yinger, J. (1995). Closed Doors, Opportunities Lost: The Continuing Costs of Housing Discrimination. Russell Sage Foundation. Retrieved from http://www.jstor.org/stable/10.7758/9781610445627

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California